AD_____

Award Number: W81XWH-07-1-0242

TITLE: Infrared Spectroscopic Imaging for Prostate Pathology Practice

PRINCIPAL INVESTIGATOR: Rohit Bhargava, Ph.D.

CONTRACTING ORGANIZATION: University of Illinois
Champaign, IL 61820

REPORT DATE: March 2009

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-03-2009	Annual	15 FEB 2008 - 14 FEB 2009

4. TITLE AND SUBTITLE

Infrared Spectroscopic Imaging for Prostate Pathology
.

5a. CONTRACT NUMBER
*H

5b. GRANT NUMBER
W81XWH-07-1-0242

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Rohit Bhargava, Ph.D.

Email:rxb@uiuc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Illinois

Champaign, IL 61820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Medical Research and Materiel Command
504 Scott Street
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT The report summarizes progress towards using Fourier transform infrared spectroscopic imaging for prostate pathology in year 2 of a 3 year award from the PCRP. The aim of the work is to enable histopathologic recognition without the use of human input or stains. The major accomplishments in the past year are: 1) A genetic algorithm based method to distinguish benign from malignant epithelium using infrared spectroscopic imaging data was shown to be effective. Large scale validation is underway. 2) A combination of IR and conventional pathology imaging has been developed. This is a critical step to potential clinical translation, and 3) A combination of IR imaging and conventional pathology shows promising results that can be explained in the context of existing practice. Larger validation studies are needed.

15. SUBJECT TERMS
Spectroscopy, prostate, histopathology, cancer, optimization, optical imaging

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	57	

Table of Contents

Page

Introduction

Prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men,[1] is a leading cause of cancer-related death and is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA)[2] or digital rectal screens are abnormal,[3] a biopsy is considered to detect or rule out cancer. Pathologic status of biopsied tissue forms the definitive diagnosis for prostate cancer and constitutes an important cornerstone of therapy and prognosis.[4] There is, hence, a need to add useful information to diagnoses and to introduce new technologies that allow efficient analyses of cancer to focus limited healthcare resources. For the reasons underlined above, there is an urgent need for high-throughput, automated and objective pathology tools. Our general hypothesis is that these requirements are satisfied through innovative spectroscopic imaging approaches that are compatible with, and add substantially to, current pathology practice. Hence, the overall aim of this project is to demonstrate the utility of novel Fourier transform infrared (FTIR) spectroscopy-based, computer-aided diagnoses for prostate cancer and develop the required microscopy and software tools to enable its application.

FTIR spectroscopic imaging is a new technique that combines the spatial specificity of optical microscopy and the biochemical content of spectroscopy.[5] As opposed to thermal infrared imaging, FTIR imaging measures the absorption properties of tissue through a spectrum consisting of (typically) 1024 to 2048 wavelength elements per pixel.[6] Since mid-IR (2-12 μm wavelength) spectra reflect the molecular composition of the tissue, image contrast arises from differences in endogenous chemical species. As opposed to visible microscopy of stained tissue that requires a human eye to detect changes, numerical computation is required to extract information from IR spectra of unstained tissue. Extracted information, based on a computer algorithm, is inherently objective and automated. Recent work has demonstrated that these determinations are also accurate and reproducible in large patient populations.[7] Hence, we focused, in the first year of this project, on demonstrating that the laboratory results could be optimized using novel approaches to fast imaging. This is a critical step, since we propose next to analyze 375 radical prostatectomy samples. We have been able to optimize data acquisition parameters and develop a novel algorithm for processing data that enables almost 50-fold faster imaging. Briefly, the idea behind the process is illustrated in Figure 1. In this performance period, we sought to acquire more data (task 1), establish the use of IR imaging for validating cancer diagnosis (task 2), develop a calibration and prediction model for grading and perform extensive validation (task 2).

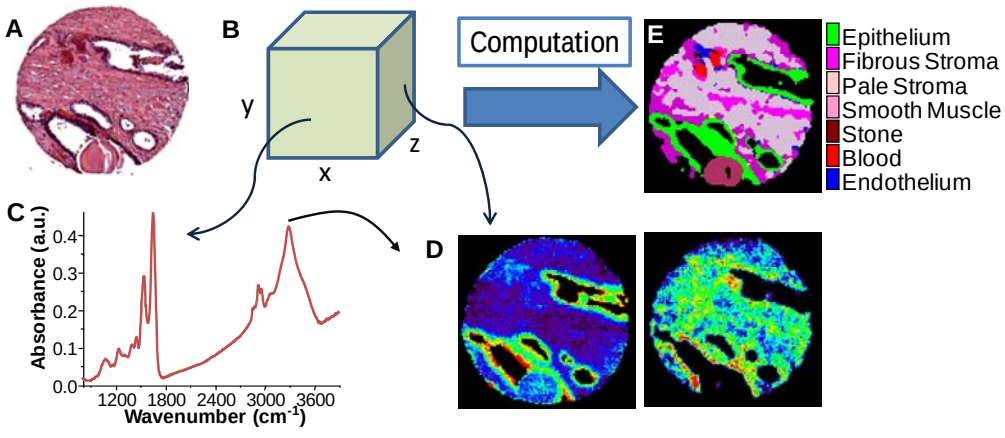


Figure 1. (A) Conventional imaging in pathology requires dyes and a human to recognize cells. In chemical imaging data cubes (B), both a spectrum at any pixel (C) and the spatial distribution of any spectral feature can be seen. e.g. in (D) nucleic acids (left, at ~1080 cm-

^{1}), and collagen specific (right, at ~ 1245 cm^{-1}) Computational tools can then convert chemical imaging data to knowledge used in pathology (E).

Body

Specific activities and tasks as per statement of work during this performance period are described below. Details of performance for the first year period are given in the past annual report which is attached for quick reference of the reviewers. :

Task 1. **Perform infrared spectroscopic imaging on prostate biopsy specimens**

Goal: Data will be acquired from samples identified in Task 2, sub-task a. 4 cm^(-1) spectral resolution data, imaging ~6 micrometer of sample per pixel will be acquired with a signal to noise ratio of greater than 1000:1. At least 375 samples will be imaged to provide as estimated 40 million spectra. Data will continuously be available for analysis in this period. (Months 8-18)

Activities: Over 5 million spectra have been acquired from approx. 475 samples using 4 cm^{-1} resolution over the 7200-720 cm^{-1} range and 6.25 micron on a side per pixel. Data handling and analysis is on-going. The data were acquired using a tissue microarray with no restrictions on age or prior PSA reading. The archiving and record keeping for such data sets became a challenge. Hence, we developed data handling tools to both maintain a database of properties as well as visualize the data in a microarray format. For example, one acquired data set is shown below in Figure 2.

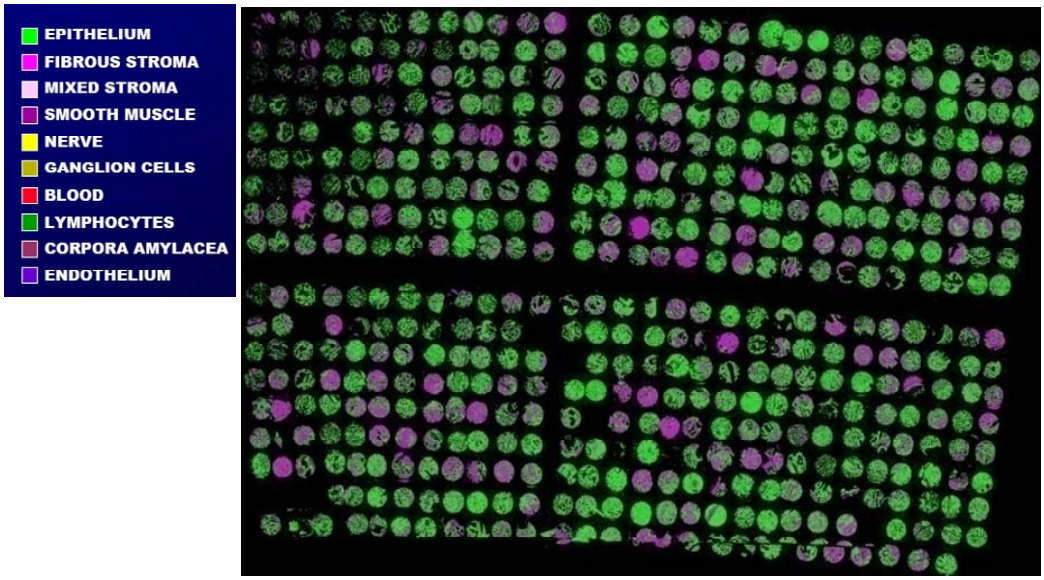


Figure 2. Approximately 475 viable samples for further analysis acquired by FT-IR imaging and classified as per optimized protocols developed previously in this project.

A second set of 460 samples were also acquired for validation studies. This large scale data acquisition has never been previously reported and is a direct result of the optimizations accomplished in year 1 of this project. Corresponding to each sample in the tissue array above, we have developed a database to store information for the patient, including age, PSA values at the time of diagnosis, Gleason grade and stage on diagnosis as well as outcome.

As per previous studies in year 1, we determined that there was a need to acquire data of a signal to noise ratio (SNR) of at least 1000:1 (or, 30 dB). One outstanding question is how to predict the required SNR for any classification task. This is a major issue in which no useful guidance

was available in the literature. In observing the data from many samples, it became clear that new tools were needed to visualize diversity and usefulness of particular samples. In particular, one key element of the protocol depends on a quality check. If contaminations exist in samples or the sample does not belong to a population that is similar to the one that was used to construct a calibration of the data, then the sample will clearly lead to incorrect results. Such a sample must be flagged during quality control but there was no obvious means to do so. Hence, we developed a new visualization system for spectrum wide analysis of the data.

First, we recall that not every point in the spectrum (Figure 1C) is actually useful in calibration or prediction. The data are reduced to a potential set of descriptors, termed metrics, which are peak height ratios, areas, positions or even spatial indices. Only a few of these metrics are useful in calibration, and consequently, in predicting histopathology. Hence, we employ the visualization only for a set of metrics. A view of the developed software and typical plot resulting from the analysis is shown in Figure 3.

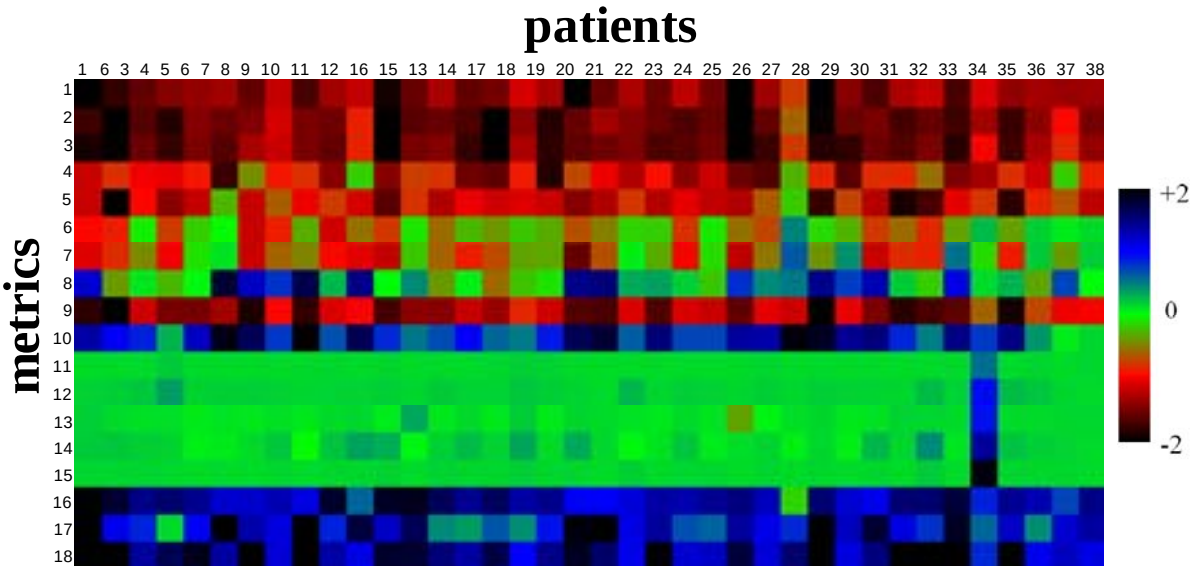


Figure 3. A Representation of metric-patient data to determine quality and consistency in large scale data analysis. Many representations are possible, including the one shown here. Here, the value of $(\mu_1-\mu_2)/\sigma$ for each metric is represented, where μ_1 is the mean of epithelium pixels for one patient for a particular metric and μ_2 is the mean of stroma pixels for one patient for a particular metric whereas σ is the standard deviation of the entire metric. Hence, $(\mu_1-\mu_2)/\sigma$ is a measure of classification potential in separating epithelium from stroma. Patient no. 34 can be seen to have outlier values that must be investigated in detail so as not to become a confounding variable.

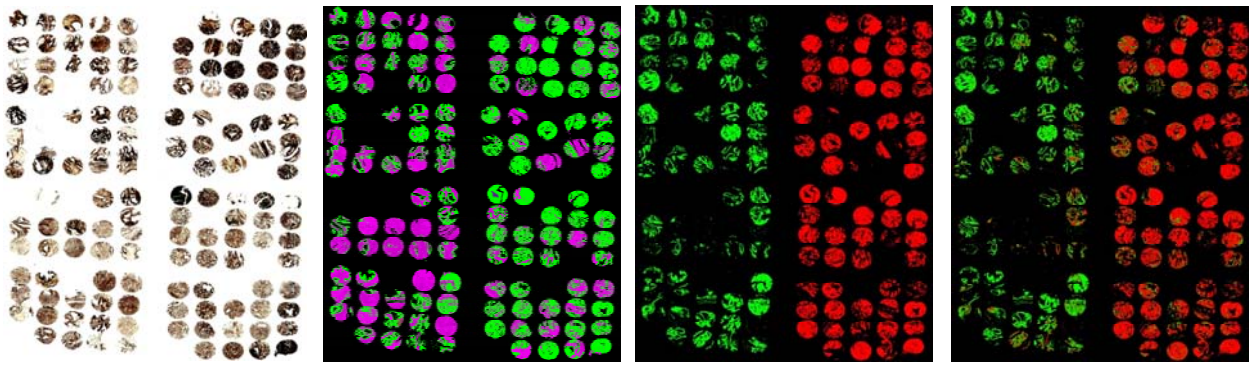
<u>Task 2. Analyze spectroscopic imaging data for biochemical markers of tumor and develop numerical algorithms for grading cancer</u>

<u>Goal</u>: Develop algorithm for malignancy recognition. Models will be constructed and optimized using Genetic Algorithms operating on identified metrics. Models will be tested and validated using ROC curves with pathologist marking as the ground truth. A protocol for segmenting benign from atypical condition will be available. (Months 11-18) Three specific aims from the statement of work (SOW) are:

a. Develop protocols and validate distinction between benign-appearing and atypical tissue (Months 12-18)
b. Develop calibration for predicting cancer grade (Months 18-22)

c. Develop protocols and validate Gleason grading of tumor (Months 18-27)

Activities: We determined metrics that were indicative of non-benign conditions using a subset of data. An example of the developed protocol applied to the training data is shown in



Array – 80 Patients | Array – Histology | Pathology Design | Pathology Result

Figure 4. A study was undertaken with 80 samples that were malignant and 80 that were benign. Some patients did not have diagnostic material and there was unrecoverable damage to other tissues yielding 140 out of 160 total specimens for classification. First, we performed histologic evaluation to determine the location of epithelial cells (figure labeled – Array Histology), which are coded in green color and stromal cells are coded in red. Stromal cells were computationally suppressed and only epithelial cells were further considered. In the prediction (Pathology Result), we were able to obtain the following accuracies : Overall pixel accuracy ~ 88.5% , 1 cancer sample classified as benign (out of 71 total cancer cases), 1 benign sample was classified as cancerous (out of a total of 69 samples classified as benign). The gold standard was a pathologist diagnosis of the samples (Pathology Design). Results show that one can potentially obtain Sensitivity and specificity exceeding human capabilities but larger validation studies are underway with other samples and confounding effects of optics need to be resolved.

Though the continued development of fast FTIR microspectroscopy**Error! Bookmark not defined.** in many laboratories worldwide represents an exciting opportunity for pathology, there is little evidence yet that the technology can add more value to the clinical enterprise than conventional pathologic examinations in prostate cancer. Hence, researchers must demonstrate both the predictive value of the technology as well as its improvement over current practice. Another intriguing question is how the new technology and existing practice can be integrated to best address needs in urology. In this manuscript, we examine one approach in which the integration of IR and H&E based information can lead to useful results. In particular, we focued on extracting morphologic measures of tissue by prior segmentation using FT-IR imaging. The extracted parameters are organized into a predictive model and evaluated for efficacy in detecting disease. The work is a first step towards integrating IR and conventional imaging for optimal use in pathology.

IR and H&E stained images were acquired from adjacent tissue samples. Although the two samples are similar, IR and stained images have different sizes, contrast mechanisms and data values. Hence, features to spatially register the images are not obvious. One option is to binarize information, but the differences in contrast mechanism may make it difficult. The second major challenge is the difference in resolution. While matching resolution is relatively straightforward, the contrast in higher resolution images makes matching details at lower resolutions difficult.

Outer shape and empty space inside the samples (lumens) are only obvious features. To overlay the two images, we first convert IR and stained images into binary images. Then, we scale up the IR image (target image) using a cubic interpolation to the spatial size of the H&E stained image (reference image). One can scale down the H&E image as well but our goal here is to use the detailed morphological information contained in the H&E stained image. In this case, most samples are elliptical or circular; thus, scale factors are determined by estimating major and minor axis of the samples. After scaling up the target image, we search for the (locally) optimal match by shifting, rotating, and scaling the target image based on a greedy algorithm. An example of the matching result is shown in Figure 5

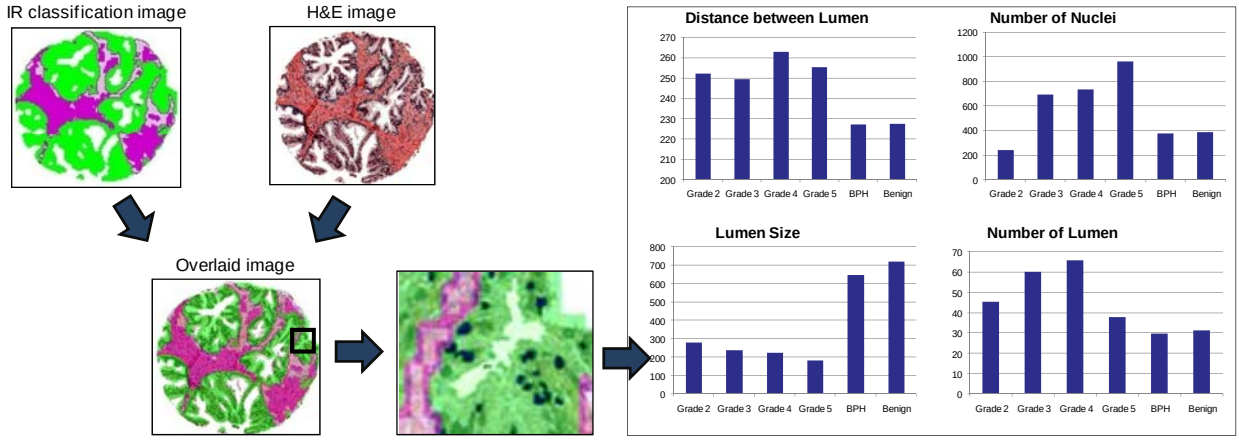


Figure 5. A synergistic blending of IR and conventional pathology (H&E) images can lead to automated extraction of spatial features that can be used for classification of tissue into benign and malignant cores. As opposed to previous efforts (e.g. figure 4), these set of data use the IR images as a guide in morphologic analyses.

A number of factors have been identified as being transformed in cancerous tissue that could potentially be used for automated analyses. One such class of factors are cellular and nuclear morphology. Properties of nuclei and lumens in normal and cancerous tissues are different but the detection and cataloging of the same is not widely practiced due to a few critical reasons. First, patient-to-patient variance and small differences in multiple clinical settings make consistent analysis of images difficult. Second, detection of epithelial nuclei may be stymied by a stromal response that is not uniform for all grades and types of cancers. We focused first on developing the methodology to obtain consistent results in the context of these challenges. We addressed two measurements: nuclear and lumen structure. The specific properties studied include the size and number of nuclei and lumens, distance from nuclei to lumens or between nuclei and lumens, and geometry of the lumens. In order to use these properties, the first step is to detect nuclei and lumens correctly from the stained images.

Lumens are elliptical, empty white spaces surrounded by epithelial cells. In normal tissues, lumens are larger in diameter and can have a variety of shapes. In cancerous tissues, lumens are progressively smaller with increasing grade and generally have less distorted shapes. Our strategy to detect lumens is to find white areas (from H&E images) whose shapes are elliptical while being located next to or within the areas where epithelial cells exist (from IR imaging data). White spots inside the samples can be found by using a proper threshold value (R, G, B > 200) but and these may include many artifacts. In our observations, artifactual lumens are relatively small and/or their shapes may be arbitrary. Hence, a simple strategy was invoked to reduce false detection. We required the size of lumens to be larger than 10 pixels and the major and minor axis ratio ($r_{major/minor}$) to be less than 3 if the size of lumens was smaller than 100

pixels. A second challenge arises from the limited amount of samples in our data set resulting in incomplete lumens for some samples. While the structures can manually be recognized to be lumens, they do not form a complete geometrical shape for easy identification. To identify these partial lumens, we first model the entire tissue core as an ellipse. The areas within the ellipse that may correspond to lumens are then restricted by two further considerations: an incomplete lumen has to have the ratio of its major to minor axis, $r_{major/minor} < 3$ and size of the lumen > 100 pixels. While these objective criteria were determined from observations of tissue structures on the array, other rules may be sought.

Nucleus detection by automated analysis is more difficult than lumen detection due to variability in staining and experimental conditions under which images were acquired. Nuclei are relatively dark and can be modeled as small elliptical areas in the stained images. The geometrical model is often confounded as multiple nuclei can be so close as to appear like one big, arbitrary-shaped nucleus. This observation illustrates both the challenge of segmenting nuclei as well as the need for high resolution imaging. Generalized detection of stained structures can prove difficult. For example, small folds or edge staining around lumens can make the darker shaded regions difficult to analyze. Here, we use the segmentation provided by the classified IR image to frame the problem. Epithelial pixels can be isolated on the H&E images using the IR overlay to provide two types of pixels: pink and blue staining, which arise from the nuclear and cytoplasmic component respectively. For nuclei restricted to epithelial cells in this manner, a set of general observations may be noted: 1) Red, Green, and Blue channel intensities are higher in nuclear pixels and lower in cytoplasmic pixels. 2) Green channel intensity is lower than other channels in both cytoplasmic and nuclear pixels. 3) In stromal cells, which are not considered here, Red channel intensity is usually higher than other channels, reflecting the pink stain. 4) A difference between Red and Blue channel intensities is small both in cytoplasmic and nuclear pixels. Based these observations, we found that converting the stained image to a new color system RG–B (|Red + Green – Blue|) could well characterize the areas where nuclei are present upon thresholding. After the color system conversion, we apply a morphological closing operator to the image to fill small holes and gaps within nuclei. The final segmentation of each individual nucleus is accomplished by using watershed algorithm. The entire segmentation process is shown
Figure **6**.

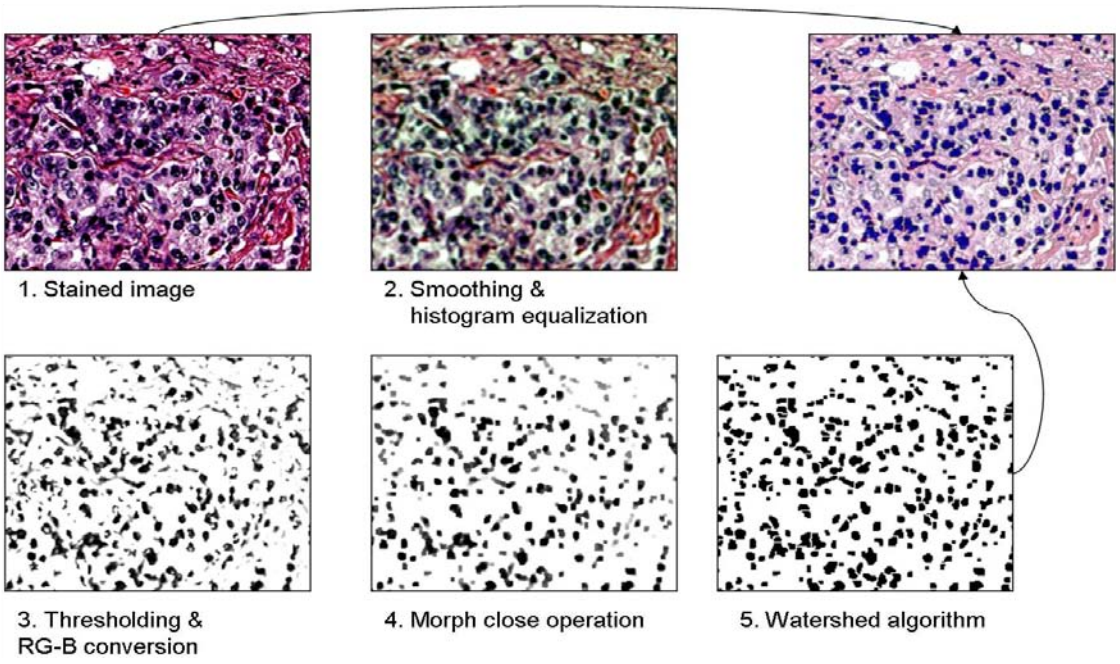


Figure 6. Step-by-step protocol for detection of nuclei. The final result of the process is a robust recognition of nuclei and correction of images for consistent morphologic analysis.

Since the raw color intensity of the stained images is variable, simple thresholds could fail to correctly segment the regions we want. To compensate for potential confusion, we adaptively determine the two threshold values. Pixels (P) where Red channel intensity is less than either of other two channels are only collected. Since we can segment epithelial cells from the IR data, pixels are assumed to be either cytoplasmic or nuclear. The threshold values become the $average(P) - \frac{2}{3} STD(P)$ for both Red(Th_{Red}) and Green(Th_{Green}) channel. It was found that the red channel intensity neither changes as much as Green channel intensity, nor is it critical to identify nuclei. The green channel intensity is skewed in cancerous tissues, however, that may increase a false discovery of nuclei in cancerous cells. To make the segmentation consistent and robust, and obtain better contrast for Green channel, adaptive histogram equalization was performed.

Following the image processing steps, we sought to use this consistent data for prediction. We developed a generative model to describe different characteristics of epithelial cells and lumens. In our model, the generative process for a tissue is 1) Create a tissue of a certain size, 2) Given a tissue size, choose areas covered by epithelial cells in a tissue, and select the number of lumens and the distance between them, 3) For each lumen, select its size, major/minor axis, and number of nuclei around it and distort it, 4) For nuclei around a lumen, select their sizes, distances to the lumen, and angle difference to the next nucleus, and place them. As generating lumens, we separate partial lumens from complete lumens since they could affect prior knowledge of the complete lumens. We assume that each lumen is independent each other and the formation of lumens is equally-likely. Thus, the probability of generating a given tissue characteristic based on a model θ is defined as:

$$P(tissue \mid \theta)$$

$$= P(S_S)P(S_E \mid S_S)P(N_L \mid S_S)P(N_{iL} \mid S_S)P(D_{Ls} \mid S_S)$$

$$\bullet \prod_{\substack{complete \\ Lumen}} \left[P(S_L)P(D_L \mid S_L)P(L_{Maj} \mid S_L)P(L_{Min} \mid S_L)P(N_{NN} \mid S_L) \prod_{\substack{Nuclei \\ around\ Lumen}} P(S_N \mid S_L)P(D_N \mid S_L)P(A_N \mid S_L) \right]$$

$$\bullet \prod \left[P(S_{iL})P(D_{iL} \mid S_{iL})P(L_{iMaj} \mid S_{iL})P(L_{iMin} \mid S_{iL})P(N_{iNN} \mid S_{iL}) \prod P(S_{iN} \mid S_{iL})P(D_{iN} \mid S_{iL})P(A_{iN} \mid S_{iL}) \right]$$

(1)

where S_S, S_E, S_L, S_{iL}, S_N are sizes of sample, epithelial cells, lumens and incomplete lumens, respectively; N_L, N_{iL}, N_{NN}, N_{iNN} are the number of complete and incomplete lumens, nuclei around a lumen, nuclei around an incomplete lumen, respectively; L_{Maj}, L_{Min} L_{iMaj}, $L_{iMin}$are lengths of major and minor axis of lumen and incomplete lumen, respectively. In addition to these geometric parameters, we also developed distortion parameters. D_L, D_{iL} are a distortion of lumen and incomplete lumen, and D_N is a distance from nuclear centers to lumen. A distortion of a lumen is defined as the distance from an ideal ellipse to the lumen on a straight line from the ideal ellipse to the center of the lumen. The ideal ellipse can be drawn by finding the major/minor axis and the center of the lumen. For the probability of a Lumen distortion, we employ a Markov chain assumption, namely, the distance from the ideal ellipse to the lumen at a certain point on the ideal ellipse is only dependent on the distance from the previous point to the lumen. A distance from a nucleus to lumen is defined in the same manner as lumen distortion.

We first proposed to use the generative model to classify the tissues samples. This is usually accomplished by computing the log-likelihood of tissues based on different classes, computing the difference between predictions based on a training set and using it as a decision function. For example, to classify a tissue into normal or cancer classes, we would first compute log-likelihood of the tissue based on both normal and cancerous samples. The difference between two log-likelihoods would give us the class to which the tissue belongs. From our study, however, we observed that just a couple of features in the log-likelihood function were determining dominantly the magnitude of the function. Regardless of the discriminating ability, selected few features dominate the decision function since likelihood is the product of probabilities. Hence, the simple measurement of more than one term that is fundamentally based on the same transition makes a larger contribution to the decision function. Since the method does not test for independently prognostic terms, the selection of features biases the likelihood values in a manner that may not be optimal for segmentation.

To resolve the issue, we employed a support vector machine (SVM) based algorithm for segmentation but based inputs to the SVM on the results from the generative model. The value of each feature is determined from the log-likelihood values obtained from the generative model as:

$$f(\text{Lumen size } = s) = \frac{1}{1+\exp\left(-\log\dfrac{P(\text{Lumen size } = s \mid \theta_{cancer})}{P(\text{Lumen size } = s \mid \theta_{normal})}\right)} \qquad \textbf{(2)}$$

The parameters for the generative model were learned from the entire dataset, and features values were calculated for SVM. The method was subsequently validated with 10-fold cross-validation. Briefly, in the 10-fold validation, a selection algorithm randomly partitions the entire dataset into 10 distinct sets, chooses 9 sets to train SVM and uses the remaining set for testing. We repeated this selection 200 times and measured four quantities: an overall accuracy, False Positive Rate (FPR), 1 – True Positive Rate (1 – TPR), and AUC (area under the ROC curve). The overall accuracy is the number of correctly classified samples over all test samples. FPR is

the number of negative samples classified as positive over all negative samples. 1 – TPR is the number of positive samples classified as negative over all positive samples. We summed all false negative and positive predictions for each 10-fold cross-validation and computed the ratios. The four quantities shown in following tables are the average of the ratios over 500 replicates.
For calculation of "test-positive" cases, cancer samples are positive samples and normal samples are negative samples. The overall accuracy is 9.27%. FPR and 1 – TPR at threshold value 0 are 8.0% and 6.8%, respectively. 1 – TPR is number of cancer samples classified as normal over all cancer samples. Accordingly, achieving lower 1 – TPR is more significant than lower FPR. In ROC curve, AUC is 0.99. All measurement are listed in Table 1 and 2.

	Average	Median	Standard Deviation	Minimum	Maximum
Accuracy (%)	92.7	92.7	1.1	90.5	95.5
AUC	0.985	0.996	0.025	0.838	1.00

Table 1. The overall accuracy and AUC of cancer and normal classification.

Threshold	Type	Average	Median	Standard Deviation	Minimum	Maximum
None	FPR (%)	8.0	7.9	1.8	3.3	13.2
	[1-TPR] (%)	6.8	7.0	1.4	3.9	11.3

Table 2. FPR and 1–TPR for cancer and normal classification

Quite interestingly, when the entire data set was used for training the classifier, an accuracy of 100% was obtained. While the observation underscores the need to be cautious in validation, it also suggests that better classification than what we have achieved may be possible with more training or better feature extraction.

In summary, of the three sub-aims in task 2, the first has been accomplished to a reasonable degree and progress on the other two is on-going (the sub-aims overlap years 2 and 3 of the project).

Key Research Accomplishments

- A genetic algorithm based method to distinguish benign from malignant epithelium using infrared spectroscopic imaging data was shown to be effective. Large scale validation is underway.
- A combination of IR and conventional pathology imaging has been developed. This is a critical step to potential clinical translation
- A combination of IR imaging and conventional pathology shows promising results that can be explained in the context of existing practice. Larger validation studies are needed.

Reportable Outcomes……

Manuscripts

Peer reviewed manuscripts published

1. R.K. Reddy, **R. Bhargava** "Automated noise reduction for accurate classification of tissue from low signal-to-noise ratio imaging data" *Anal. Chem.*, Under Review (2009)
2. X. Llora, A.Priya, **R. Bhargava** "Observer-Invariant Histopathology using Genetics-Based Machine Learning" *Nat. Computing*, 8, 101-120 (2009)

Book Chapters

1. R.K. Reddy, **R. Bhargava** "Chemometric methods for biomedical vibrational spectroscopy and imaging", P. Matousek and M.D. Morris, eds. (2009 - Anticipated)
2. A.K. Kodali, **R. Bhargava** "Nanostructured Probes to Enhance Optical and Vibrational Spectroscopic Imaging for Biomedical Applications", Y.Y. Fu and A. Narlikar, eds. (2009 - Anticipated)
3. **R. Bhargava**, I.W. Levin "Prostate Cancer Diagnosis by FTIR Imaging", M. Diem, P.R. Griffiths and J. Chalmers, eds (2008)

Published abstracts

1. RK Reddy, **R Bhargava** "Automated and fast histologic characterization in urology: progress towards an unmet clinical need", Urology: Diagnostics, Therapeutics, Robotics, Minimally Invasive, and Photodynamic Therapy, BiOS 2009, San Jose, CA, In press
2. R.K. Reddy, F.N. Pounder, **R. Bhargava** "Validating the cancer diagnosis potential of mid-infrared spectroscopic imaging", SPIE Photonics West - BiOS 2009, San Jose, CA, In press
3. J. Ip, **R. Bhargava** "Integrating instrumentation, computation and sampling for a high throughput approach to automated histology by mid-infrared microscopy", Advanced Biomedical and Clinical Diagnostic Systems VII, SPIE Photonics West - BiOS 2009, San Jose, CA, In press
4. M.J. Walsh, F.N. Pounder, **R. Bhargava** "Spectral pathology in breast cancer using mid-infrared spectroscopic imaging", Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues VII, SPIE Photonics West - BiOS 2009, San Jose, CA , In press

Presentations

Invited conference presentations

First author is the presenting author; First author is also the invited author unless indicated by *

1. **R. Bhargava** "Imaging: Does it really offer more than 'just' pretty pictures", SAS 50 years symposium, Pittcon 09, Chicago, March 2009
2. **R. Bhargava** "The critical role of controlled quality of spectral information and sampling on automated histologic recognition", Pittcon 09, Chicago, March 2009
3. **R. Bhargava** F.N. Pounder, X. Llora and R.K. Reddy "Enhancing the tissue segmentation capability of fast infrared spectroscopic imaging via chemometric methods", FACSS08, Reno, September 2008
4. **R. Bhargava**, F.N. Keith, R.K. Reddy and A.K. Kodali "Practical infrared spectroscopic imaging instrumentation for translating laboratory results to clinical settings", FACSS08, Reno, September 2008
5. **R. Bhargava** "Spectroscopic Imaging for an Automated Approach to Histopathologic Recognition in Prostate Tissue" 82nd Annual North Central Section American Urological Association Meeting, Chicago, September 2008

6. **R. Bhargava**, R.K. Reddy, A.K. Kodali "Ultrafast mid-infrared spectroscopic imaging by combined computational and experimental optimizations" ISSSR 2008, Hoboken, June 2008

Other invited presentations

1. Department of Chemistry, University of Kentucky, Knoxville, 2009
2. BioInterest Group Seminar, Mechanical Science and Engineering, UIUC, 2008
3. Lester Wolfe Workshop, MIT, 2008
4. Translational Biomedical Research Seminar, Veterinary Medicine, UIUC, 2008
5. Vistakon, A Division of Johnson and Johnson, Jacksonville, 2008

Contributed presentations

*First author is the presenting author, unless indicated by **

1. RK Reddy, **R Bhargava** "Automated and fast histologic characterization in urology: progress towards an unmet clinical need", Urology: Diagnostics, Therapeutics, Robotics, Minimally Invasive, and Photodynamic Therapy, BiOS 2009, San Jose, CA
2. R.K. Reddy, F.N. Pounder, **R. Bhargava** "Validating the cancer diagnosis potential of mid-infrared spectroscopic imaging", SPIE Photonics West - BiOS 2009, San Jose, CA
3. J. Ip, **R. Bhargava** "Integrating instrumentation, computation and sampling for a high throughput approach to automated histology by mid-infrared microscopy", Advanced Biomedical and Clinical Diagnostic Systems VII, SPIE Photonics West - BiOS 2009, San Jose, CA
4. M.J. Walsh, F.N. Pounder, **R. Bhargava** "Spectral pathology in breast cancer using mid-infrared spectroscopic imaging", Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues VII, SPIE Photonics West - BiOS 2009, San Jose, CA
5. R. Bhargava, A.K. Kodali, F.N. Pounder, R.K. Reddy "High-speed Infrared Spectroscopic Imaging for Tissue Histopathology", EAS 2008, Somerset, November 2008
6. R.K. Reddy, **R. Bhargava** "Robustness of Histological Recognition in Tissues Using Fourier Transform Infrared Spectroscopic Imaging" FACSS 2008, Reno, October 2008

Infomatics such as databases

Databases of spectra and spectral data sets have been combined to update a website formed during the first year of this project –metaspectra.org

Funding applied for based on work supported by this award

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:
Infrared spectroscopic imaging for a systems approach to prostate pathology (Role: PI)
Source of Support: Charlotte Geyer Foundation
Total Award Amount: $100,000 Total Award Period Covered: 1/1/2009-12/31/2009
Location of Project: Urbana, IL
Person-Months Per Year Committed to the Project. Cal: 1.0 Acad: Sumr:

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:
Nanofilter-based Infrared Spectroscopic Imaging (Role: PI)
Source of Support: Grainger Foundation
Total Award Amount: $100,000 Total Award Period Covered: 12/1/2008-11/30/2009
Location of Project: Urbana, IL

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:
Nanofilters for prostate pathology using infrared spectroscopic imaging (Role: PI)
Source of Support: National Cancer Institute Center for Nanotechnology Excellence
Total Award Amount: $198,800 Total Award Period Covered: 01/01/2008-8/31/2009
Location of Project: Urbana, IL
Person-Months Per Year Committed to the Project. Cal: 0.5 Acad: Sumr:

Support: ☐ Current	☒ Pending	☐ Submission Planned in Near Future	☐ *Transfer of Support	
Project/Proposal Title:				
CDI-Type I: Chemical Imaging: From Data to Knowledge (Role: PI) – Preliminary proposal				
Source of Support: National Science Foundation				
Total Award Amount: $ 640,272	Total Award Period Covered: 06/01/2009-5/30/2012			
Location of Project: Urbana, IL				
Person-Months Per Year Committed to the Project.		Cal:	Acad:	Sumr: 1.0
Support: ☐ Current	☒ Pending	☐ Submission Planned in Near Future	☐ *Transfer of Support	
Project/Proposal Title:				
CDI-Type II: combinatorial optimization framework for analysis and design of systems with multiple interacting elements (Role: co-PI) – Preliminary proposal ***[Selected for full proposal]***				
Source of Support: National Science Foundation				
Total Award Amount: $ 1 464,540	Total Award Period Covered: 06/01/2009-5/30/2012			
Location of Project: Urbana, IL				
Person-Months Per Year Committed to the Project.		Cal:	Acad:	Sumr: 0.5
Support: ☐ Current	☒ Pending	☐ Submission Planned in Near Future	☐ *Transfer of Support	
Project/Proposal Title:				
Development of practical mid-infrared spectroscopic imaging technology for cancer pathology (Role: PI)				
Source of Support: National Institutes of Health				
Total Award Amount: $ 1 879, 397	Total Award Period Covered: 06/01/2009-5/30/2014			
Location of Project: Urbana, IL				
Person-Months Per Year Committed to the Project.		Cal:	Acad:	Sumr: 1.0
Support: ☐ Current	☒ Pending	☐ Submission Planned in Near Future	☐ *Transfer of Support	
Project/Proposal Title:				
Infrared microscopy for a systems approach to prostate pathology (Role: PI)				
Source of Support: National Institutes of Health				
Total Award Amount: $ 1 832, 819	Total Award Period Covered: 08/01/2009-7/30/2014			
Location of Project: Urbana, IL				
Person-Months Per Year Committed to the Project.		Cal:	Acad:	Sumr: 1.0

Employment or research opportunities applied for and/or received based on experience/training supported by this award.

Dr. Gokulakrishnan Srinivasan, a post-doctoral fellow working on this project obtained employment with Bruker Optics.

Conclusion……

The work accomplished demonstrates clear potential and preliminary protocols for classifying prostate tissue. If the protocols are validated in on-going larger studies, a new tool for prostate histopathology will be available.

So What Section

If the reported progress is sustained, an automated method for prostate pathology will be available that can rapidly determine the presence of cancer in biopsies and aid pathologists in making accurate decisions.

References……

[1] A Jemal, R Siegel, E Ward, T Murray, J Xu, C Smigal, MJ Thun Cancer statistics, 2006 *CA Cancer J Clin* **56**, 106-130 (2006).

[2] SM Gilbert, CB Cavallo, H Kahane, FC Lowe Evidence suggesting PSA cutpoint of 2.5 ng/mL for prompting prostate biopsy: Review of 36,316 biopsies. *Urology* **65**, 549-553 (2005).

[3] PF Pinsky, GL Andriole, BS Kramer, RB Hayes, PC Prorok, JK Gohagan, Prostate, Lung, Colorectal and Ovarian Project Team Prostate Biopsy Following a Positive Screen in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial *J Urol* **173**, 746-750 (2005). discussion 750-751.

[4] PA Humphrey *Prostate Pathology* American Society of Clinical Pathology, Chicago (2003).

[5] EN Lewis, PJ Treado, RC Reeder, GM Story, AE Dowrey, C Marcott, IW Levin Fourier transform spectroscopic imaging using an infrared focal-plane array detector *Anal. Chem.* **67**, 3377-3384 (1995).

[6] R Bhargava, SQ Wang, JL Koenig Processing FTIR Imaging Data for Morphology Visualization *Appl Spectrosc* **54**, 1690-1706 (2000).

[7] DC Fernandez, R Bhargava, SM Hewitt, IW Levin Infrared spectroscopic imaging for histopathologic recognition *Nat. Biotechnol.* **23**, 469-474 (2005).

Nat Comput (2009) 8:101 120
DOI 10.1007/s11047 007 9056 6

Observer-invariant histopathology using genetics-based machine learning

Xavier Llorà · Anusha Priya · Rohit Bhargava

Published online: 11 October 2007

Abstract Prostate cancer accounts for one third of noncutaneous cancers diagnosed in US men and is a leading cause of cancer related death. Advances in Fourier transform infrared spectroscopic imaging now provide very large data sets describing both the structural and local chemical properties of cells within prostate tissue. Uniting spectroscopic imaging data and computer aided diagnoses (CADx), our long term goal is to provide a new approach to pathology by automating the recognition of cancer in complex tissue. The first step toward the creation of such CADx tools requires mechanisms for automatically learning to classify tissue types a key step on the diagnosis process. Here we demonstrate that genetics based machine learning (GBML) can be used to approach such a problem. However, to efficiently analyze this problem there is a need to develop efficient and scalable GBML implementations that are able to process very large data sets. In this paper, we propose and validate an efficient GBML technique NAX based on an incremental genetics based rule learner. NAX exploits mas sive parallelisms via the message passing interface (MPI) and efficient rule matching using hardware implemented operations. Results demonstrate that NAX is capable of performing prostate tissue classification efficiently, making a compelling case for using GBML implementations as efficient and powerful tools for biomedical image processing.

X. Llorà (✉)
National Center for Supercomputing Applications, University of Illinois at Urbana Champaign,
1205 W. Clark Street, Urbana, IL 61801, USA
e mail: xllora@uiuc.edu

A. Priya R. Bhargava
Department of Bioengineering, University of Illinois at Urbana Champaign, 1304 W. Springfield Ave.,
Urbana, IL 61801, USA

A. Priya
e mail: priya@uiuc.edu

R. Bhargava
e mail: rxb@uiuc.edu

R. Bhargava
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana Champaign,
405 N. Mathews Ave., Urbana, IL 61801, USA

Keywords Observer invariant histopathology · Genetics based machine learning · Learning Classifier Systems · Hardware acceleration · Vector instruction · SSE2 · MPI · Massive parallelism

1 Introduction

Pathologist opinion of structures in stained tissue is the definitive diagnosis for almost all cancers and provides critical input for therapy. In particular, prostate cancer accounts for one third of noncutaneous cancers diagnosed in US men. Hence, it is, appropriately, the subject of heightened public awareness and widespread screening. If prostate specific antigen (PSA) or digital rectal screens are abnormal, a biopsy is needed to definitively detect or rule out cancer. Pathologic status of biopsied tissue not only forms the definitive diagnosis but constitutes an important cornerstone of therapy and prognosis. There is, however, a need to add useful information to diagnoses and to introduce new technologies that allow economical cancer detection to focus limited healthcare resources. In pathology practice, widespread screening results in a large workload of biopsied men, in turn, placing a increasing demand on services. Operator fatigue is well documented and guidelines limit the workload and rate of examination of samples by a single operator. Importantly, newly detected cancers are increasingly moderate grade tumors in which pathologist opinion variation complicates decision making.

For the reasons above, there is an urgent need for automated and objective pathology tools. We have sought to address these requirements through novel Fourier transform infrared (FTIR) spectroscopy based, computer aided diagnoses for prostate cancer and develop the required microscopy and software tools to enable its application. FTIR spectroscopic imaging is a new technique that combines the spatial specificity of optical microscopy and the biochemical content of spectroscopy. As opposed to thermal infrared imaging, FTIR imaging measures the absorption properties of tissue through a spectrum consisting of (typically) 1024 2048 wavelength elements per pixel. Since IR spectra reflect the molecular composition of the tissue, image contrast arises from differences in endogenous chemical species. As opposed to visible microscopy of stained tissue that requires a human eye to detect changes, numerical computation is required to extract information from IR spectra of unstained tissue. Extracted information, based on a com puter algorithm, is inherently objective and automated (Lattouf and Saad 2002; Fernandez et al. 2005; Levin and Bhargava 2005; Bhargava et al. 2006).

Uniting spectroscopic imaging data and computer aided diagnoses (CADx), we seek to provide a new approach to pathology by automating the recognition of cancer in complex tissue. This is an exciting paradigm in which disease diagnoses are objective and repro ducible; yet do not require any specialized reagents or human intervention. The first step toward the creation of such CADx tools requires mechanisms for reliable and automated tissue type classification. In this paper we demonstrate how genetics based machine learning tools can achieve such a goal. Interpretability of the learned models and efficient processing of very large data sets have lead us to rule based models easy to interpret and genetics based machine learning inherent massively parallel methods with the required scalability properties to address very large data sets. We present the method and the efficiency enhancement techniques proposed to address automated tissues classifica tion. When pushed beyond the relatively small problems traditionally used to test such methods, an need for efficient and scalable implementations becomes a key research topic

that needs to be addressed. We designed the proposed a technique with such constraints in mind. A modified version of an incremental genetics based rule learner that exploits massive parallelisms via the message passing interface (MPI) and efficient rule matching using hardware oriented operations. We name this system NAX. NAX is compared to traditional and genetics based machine learning techniques on an array of publicly available data sets. We also report the initial results achieved using the proposed technique when classifying prostate tissue.

The remainder of the paper is structured as follows. We present an overview of the problem addressed in Sect. 2, paying special attention to tissue classification. We discuss in Sect. 3 the hurdles that traditional genetics based machine learning implementations face when applied to very large data sets. Section 4 presents our solution to those hurdles. We also describe the incremental rule learner proposed for tissue classification. Last, we summarize results on publicly available data sets and the preliminary results for tissue classification on a prostate tissue microarray in Sect. 5. Finally, in Sect. 6, we present conclusions and further work.

2 Biomedical imaging and data mining

This section presents an overview of the problem addressed in this paper. We first intro duce infrared spectroscopic imaging as a potentially powerful tool for cancer diagnosis and prognosis. Then, we explore the protocols that provide raw high quality data that for data mining. Finally, we conclude by focusing on the key task, tissue classification, by focusing on prostate tissue.

2.1 Infrared spectroscopy and imaging for cancer diagnosis and prognosis

Infrared spectroscopy is a well established molecular technique and is widely used in chemical analyses. The fundamental principle governing the response of any material is that the vibrational modes of molecules are resonant in energy with photons in the mid infrared region (2 14 mm) of the electromagnetic spectrum. Hence, when photons of energy that are resonant with the material's molecular composition are incident, a number are absorbed. The number absorbed is directly proportion to the number of chemical species that are excited. Hence, any material has a characteristic frequency dependent absorption profile called a spectrum. An infrared spectrum is often termed the "optical fingerprint" of a material as it can help uniquely identify molecular composition see Fig. 1.

Researchers, including us, have contributed to develop an imaging version of spec troscopy that is essentially similar to an optical microscope. In this mode of spectroscopy, images are acquired in the manner of optical microscopy with one important difference. Instead of measuring the intensity of three colors for a visible image, several thousand intensity values are acquired at each pixel in the image as a function of wavelength (spectrum at each pixel). The resulting data set is three dimensional (2 spatial and 1 spectral indices) consisting typically of a size $256 \times 256 \times 1024$, but extending to sizes such as $3500 \times 3500 \times 2048$. Since each data point is stored as a 16 bit number, the data size typically runs into several tens to hundreds of gigabytes.

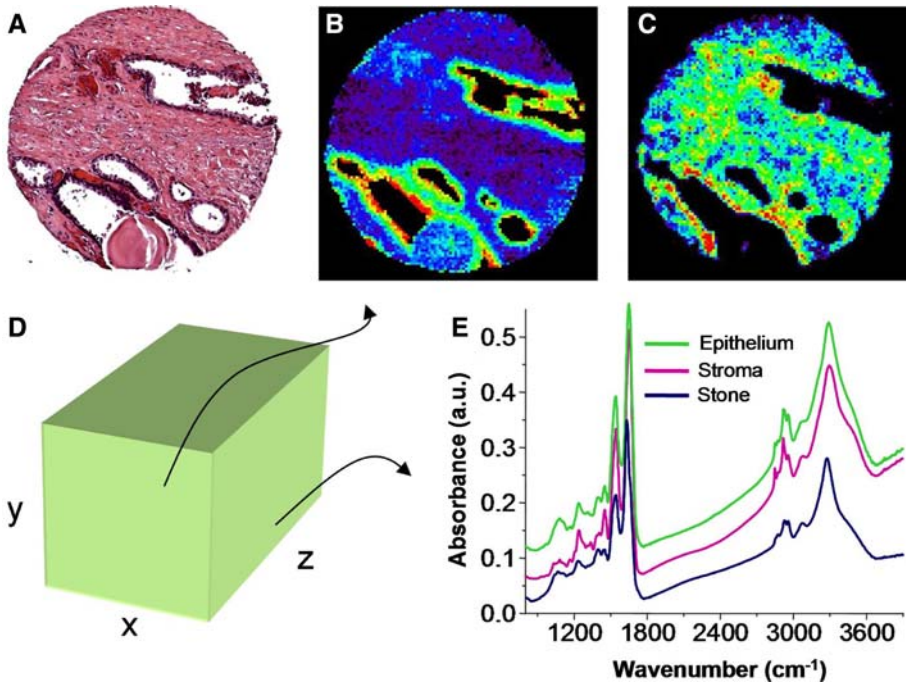


Fig. 1 Conventional staining and automated recognition by chemical imaging. (**A**) Typical H&E stained sample, in which structures are deduced from experience by a human. Highlights of specific regions in the manner of H&E is possible using FTIR imaging without stains. (**B**) Absorption at 1080 cm ^{1} commonly attributed to nucleic acids and (**C**) to proteins of the stroma. The data obtained is 3 dimensional (**D**) from which spectra (**E**) or images at specific spectral features may be plotted

2.2 Mining the spectra: Two sequential problems

Though the continued development of fast FTIR microspectroscopy represents an exciting opportunity for pathology, handling the resultant data and rapidly providing classifications remains a critical challenge. First, the sheer volume of data potentially larger than 10 GB a day represents an organizational and retrieval challenge. Next, extraction of useful information in short time periods requires the formulation of optimal protocols. Third, the automated cancer segmentation problem is very complex and offers a number of routes and levels of data that need to be analyzed to determine the optimal approach for application in a laboratory.

The typical application is the need to extract information from the data set such that it is clinically relevant. Hence, the output of the data mining algorithm to be developed is well bounded and clearly defined. For example, in the prostate there are two levels of interest. In the first level, the pathologist examines the tissue to determine if there are any epithelial cells. Since more than 95% of prostate cancers arise in epithelial cells, transformations in this class of cells forms the diagnostic basis and a primary determinant of therapy. Other cell types of interest are lymphocytes that may indicate inflammation, blood vessel density that may indicate the development of new blood supply indicative of cancer growth and nerves that may be invaded by cancer cells. Hence, any automated approach to pathology must first identify cell types accurately. The second step in pathology follows. Once

epithelial cells are located, their spatial patterns are indicative of disease states. In our imaging approach, we can identify both spatial patterns as well as chemical patterns in epithelial cells. Hence, the task would be to use either or both to classify disease. In this paper, we focus only on the accurate identification/classification of tissue types as the first step of the path that leads to obtaining the correct pixels of epithelium.

2.3 Tissue classification for prostate arrays

Prostate tissue is structurally complex, consisting primarily of glandular ducts lined by epithelial cells and supported by heterogeneous stroma. This tissue also contains blood vessels, blood, nerves, ganglion cells, lymphocytes and stones (which are comprised of luminal secretions of cellular debris) that organize into structure measuring from tens to hundreds of microns. These structures are readily observable within stained tissue using bright field microscopy at low to medium magnifications. Hence, in applying FTIR imaging (Levin and Bhargava 2005), we obtain the common structural detail employed clinically and, additionally, spectral information indicative of tissue biochemistry. As histologic classes contain identical chemical components, infrared vibrational spectra are similar but reveal small differences in specific absorbance features. The technique pro posed by Fernandez et al. (2005) examines each cell types' spectra and transforms each spectrum into a vector of describing features usually around the hundreds. A complete description of this process is beyond the scope of this paper and can be found elsewhere (Fernandez et al. 2005). Each pixel (cell present in the slice of micro array under analysis) has an assigned spatial position in the array while the tissue type is assigned by a highly experienced pathologist. Thus, the tissue classification can be cast into a supervised classification problem (Mitchell 1997), where all the attributes are real valued and the class is the tissue type ten classes: *ephithelium*, *fibrous stroma*, *mixed stroma*, *muscle*, *stone*, *lymphocytes*, *endothelium*, *nerve*, *ganglion*, and *blood*. Figure 2 presents tissue types that can be assigned by examining a stained image obtained, after the FTIR microsprectroscopy on unstained tissue,by the pathologist. Each marked pixel in the image becomes a train ing example; hence, the usual smallest data set is around hundreds of thousand records per array.

3 Larger, bigger, and faster genetics-based machine learning

Bernadó et al. (2001) presented a first empirical comparison between genetics based machine learning techniques (GBML) and traditional machine learning approached. The authors reported that GBML techniques were as competent as traditional techniques. Later, Bacardit and Butz (2006) repeated the analysis, obtaining similar results. Most of the experiments presented on both papers used publicly available data sets provided by the *University of California at Irvine* repository (Merz and Murphy 1998). Most of the data sets are defined over tens of features and up to few thousands of records in the larger cases. However, a key property of GBML approaches is its intrinsic massive parallelism and scalability properties. Cantú Paz (2000) presented how efficient and accurate genetics algorithms could be assembled, and Llorà (2002) presented how such algorithms can be efficiently used for machine learning and data mining. However, there are elements that need to be revisited when we want to efficiently apply GBML techniques to large data sets such as the one described in the previous section.

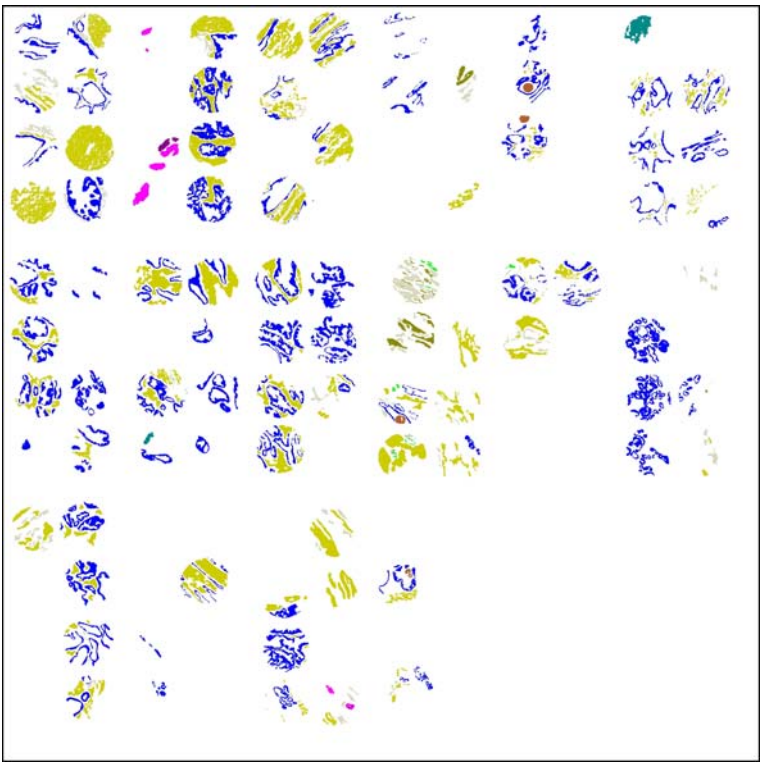


Fig. 2 The figure presents the tissue labeling provided by a pathologist biopsy section of human prostate tissue. Each spot represents the section of a needle. Different colors represent different tissue types

The GBML techniques require evaluating candidate solutions against the original data set matching the candidate solutions (e.g., rules, decision trees, prototypes) against all the instances in the data set. Regardless of the flavor used, Llorà and Sastry (2006) showed that, as the problem grows, rule matching governs the execution time. For small data sets (teens of attributes and few thousands of records) the matching process takes more than 85% of the overall execution time marginalizing the contribution of the other genetic operators. This number increases to 98% and above, when we move to data sets with few hundreds of attributes and few hundred thousands of records. More than 98% of the time is spent evaluating candidate solutions. Each evaluation can be computed in parallel. Moreover, the evaluation process may also be parallelized on very large data sets by splitting and distributing the data across the computational resources. A detailed description of the parallelization alternatives of GBML techniques can be found else where (Llorà 2002).

Currently available off the shelf GBML methods and software distributions (Barry and Drugow itsch 1997; Llorà 2006) do not usually target large data sets. The two main bottlenecks are large memory footprints and sequential processing oriented processes. Generally speaking, they were designed to run on single processor machines with enough memory to fit the entire data set. Hence, designers did not paying much

attention to the memory footprint required to store the data set usually completely loaded into memory and the population of candidate solutions. These large complex structures were geared to facilitate the programming effort, but they are not designed toward the efficient evaluation of the candidate solutions. However, efforts have been made to push GBML methods into domains which require processing large data sets. Three different works need to be mentioned here. Flockhart (1995) proposed and implemented GA MINER, one of the earliest effort to create data mining systems based on GBML systems that scale across symmetric multi processors and massively parallel multi processors. Flockhart (1995) reviewed different encoding and parallelization schemes and conducted proper scalability studies. Llorà (2002) explored how fine grained parallel genetic algorithms could become efficient models for data mining. Theoretical analysis of performance and scalability were developed and validated with proper simulations. Recently, Llorà and Sastry (2006) explored how current hardware can efficiently speed up rule matching against large data sets. These three approaches are the basis of the incremental rule learning proposed in the next section to approach very large data sets.

Another important issue in real world problems is the class distribution. Usually most real problems have a clear class imbalance. Recently, Orriols Puig and Bernadó Mansilla (2006) have revisited this issue, showing how GBML techniques successfully learn and maintain proper descriptions for those minority classes. If not designed properly, descriptions of majority classes will tend to govern the learned models, starving the description of minority classes. Prostate tissue classification is a clear example of extreme class imbalance. Figure 3 presents the tissue type class distribution. The smaller tissue type has 64 records, where as the larger classes have several tens of thousands records. hence, the developed approaches must account for class size variation.

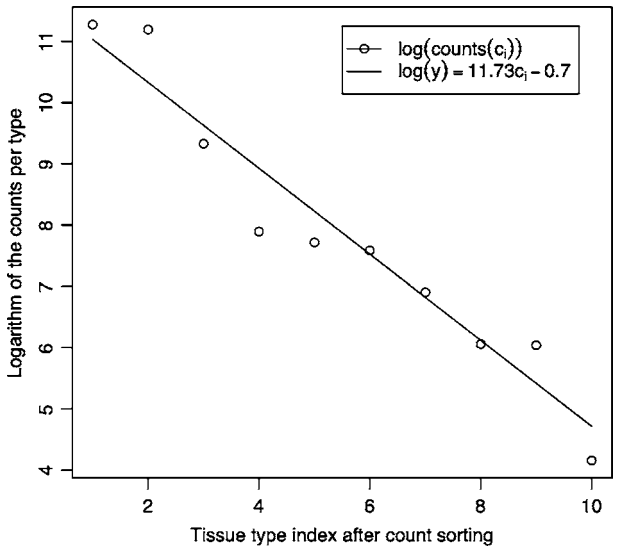


Fig. 3 Figure shows the tissue class distribution. Once the classes are reordered according to their frequency in the data set, we can easily appreciate the extreme imbalance the smaller tissue type has 64 records, where as the larger classes have several tens of thousands records

4 The road to tractability

We describe in this section the steps we took to design a GBML method (NAX) able to deal with very large data sets with class imbalance. NAX evolves, one at a time, maximally general and maximally accurate rules. Then, the covered instance are removed and another maximally general and maximally general rule is evolved and added to the previously stored one forming a decision list. This process continues until no uncovered instances are left this process is also referred as the sequential covering procedure (Cordón et al. 2001). Llorà et al. (2005) showed that maximally general and maximally accurate rules (Wilson 1995) could also be evolved using Pittsburgh style Learning Classifier Systems. Later, Llorà et al. (2007) showed that competent genetic algorithms (Goldberg 2002) evolve such rules quickly, reliably, and accurately. The rest of this section describes (1) efficient implementation techniques to deal with very large data sets, (2) the impact of class imbalance, and (3) the NAX algorithm proposed.

4.1 Efficient implementations

As introduced earlier, when dealing with very large data sets, and regardless of the flavor of the GBML technique used, we may spend up to 98% of the computational cycles trying to match rules to the original data set (Llorà and Sastry 2006). Each solution evaluation is independent of each other and, hence, it can be computed in parallel. Moreover, even the matching nature of a rule the representation we will use from now on is highly parallel, since conditions require performing simultaneous checks against different attributes per record. Thus, efficient implementation can take advantage of parallelizing both elements.

4.1.1 Exploiting the hardware acceleration

Recently, multimedia and scientific applications have pushed CPU manufactures to include support for vector instructions again in their processors. Both applications areas require heavy calculations based on vector arithmetic. Simple vector operations such as *add* or *product* are repeated over and over. During 1980s and 1990s supercomputers, such as Cray machines, were able to issue hardware instructions that enabled basic vector arithmetics. A more constrained scheme, however, has made its way into general purpose processors thanks to the push of multimedia and scientific applications. Main chip manufactures IBM, Intel, and AMD have introduced vector instruction sets Altivec, SSE3, and 3DNow$^+$ that allow vector operations over packs of 128 bits by hardware. We will focus on a subset of instructions that are able to deal with floating point vectors. This subset of instructions manipulate groups of four floating point numbers. These instructions are the basis of the fast rule matching mechanism proposed.

Our goal is to evolve a set of rules that correctly classifies the current data set rom prostate tissue. Using a knowledge representation based on rules allows us to inspect the learned model, gaining insight into the biological problem as well. All the attributes of the domain are real value and the conditions of the rules need to be able to express conditions in a $\Re^n$ spaces. We use a similar rule encoding to the one proposed by Wilson (2000b) a variation of the original work proposed by Wilson (2000a) and later reviewed by Stone and Bull (2003) and widely used in the GBML community. Rules express the conjunction of tests across attributes. Each test may be defined in multiple flavors but, without loss of

generality, we picked a simple interval based one. A simple example of an *if then* rule, could be expressed as follows:

$$1.0 \leq a_0 \leq 2.3 \wedge \cdots \wedge 10.0 \leq a_n \leq 23 \rightarrow c_1 \quad (1)$$

Where the condition is the conjunction of the different attribute tests and the outcome is the predicted class a tissue type. We also allow a special condition `don't care` which just always returns `true` , allowing condition generalization. The rule below illustrates an example of a generalized rule.

$$1.0 \leq a_0 \leq 2.3 \wedge -3.0 \leq a_3 \leq 2 \longrightarrow c_1 \quad (2)$$

All attributes except a_0 and a_3 were marked as `don't care`.

Each condition can be encoded using 2 floating point numbers per condition, where α_i contains the lower bound of the condition and ω_i its upper bound. Thus, the condition $\alpha_i \leq a_0 \leq \omega_i$ just requires to store the two floating point numbers. For efficiency reasons we store them in two separate vectors, on containing the lower bounds and the other con taining the upper bounds. The position in a vector indicates the attribute being tested. The `don't care` condition is simply encoded as $\alpha_i > \omega_i$ and, hence, we do not need to store any extra information.

Matching a rule requires performing the individual condition tests before the final *and* operation can be computed. Vector instruction sets improve the performance of this pro cess by performing four operations at once. Actually, this process may be regarded as four parallel running pipelines. The process can be further improved by stopping the matching process when one test fails since that will turn the condition into false.

Figure 4 presents a `C` implementation the proposed hardware supported rule matching. The code assumes that the two vectors containing the upper and lower bounds are provided and records are stored in a two dimensional matrix. Figure 5 presents the vectorized implementation of the code presented in Fig. 4 using SSE2 instructions. Exploiting the hardware available can speed between 3 and 3.5 times the matching process, as also shown elsewhere (Llorà and Sastry 2006).

4.1.2 Massive parallelism

Since most of the time is spent on the evaluation of candidate rules when dealing with large data sets, our next goal was to find a parallelization model that could take advantage of this peculiarity. Due the quasi embarrassing parallel (Grama et al. 2003) nature of the candi date rule evaluation, we designed a coarse grain parallel model for distributing the evaluation load. Cantú Paz (2000) proposed several schemes, showing the importance of the trade off between computation time and time spent communicating. When designing the parallel model, we focused on minimizing the communication cost. Usually, a feasible solution could be a master/slave one the computation time is much larger than the communication time. However, GBML approaches tend to use rather large populations, forcing us to send rule sets to the evaluation slaves and collect the resulting fitness. These schemes also increment the sequential sections that cannot be parallelized, threatening the overall speedup of the parallel implementation as a result of Ambdhals law (Amdahl 1967).

To minimize such communication cost, each processor runs an identical `NAX` algorithm. They are all seeded in the same manner, hence, performing the same genetic operations and only differing in the portion of the population being evaluated. Thus, the population is

```
 1. void match_seq_rule_set ( RuleSet * rs, InstanceSet is, int iDim, int iRows ) {
 2.    int i,j,k,iCnt,iClsIdx,iGround,iPred;
 3.    register int iMatcheable;
 4.    Instance ins;
 5.
 6.    iClsIdx = rs->iCorrectedDim;
 7.    clean_fitness_rules_set(rs);
 8.    for ( i=0 ; i<iRows ; i++ ) {
 9.       ins = is[i];
10.       iPred=-1;
11.       for ( j=0 ; iPred==-1 && j<rs->iLen ; j++ ) {
12.          iMatcheable = 1;
13.          for ( iCnt=0,k=j*(rs->iCorrectedDim+VBSIF) ;
14.                iMatcheable && k<j*(rs->iCorrectedDim+VBSIF)+rs->iDim ;
15.                k++,iCnt++ ) {
16.             iMatcheable = iMatcheable &&
17.                           !( (rs->pfLB[k]<=rs->pfUB[k]) &&
18.                           ( ins[iCnt]<rs->pfLB[k] || ins[iCnt]>rs->pfUB[k]));
19.          }
20.          if ( iMatcheable )
21.             iPred = rs->pfLB[j*(rs->iCorrectedDim+VBSIF)+rs->iCorrectedDim];
22.       }
23.       iPred = (iPred==-1)?rs->iClasses:iPred;
24.       iGround=(int)ins[iClsIdx];
25.       rs->pConfMat[iGround][iPred]++;
26.    }
27. }
```

Fig. 4 This figure presents a sequential implementation of the rule matched process in `C` . A rule set is match against a data set. Lines 16, 17, and 18 implement the condition test for one attribute. The implementation also computes the confusion matrix that contains the ground truth versus predicted class

treated as collection of chunks where each processor evaluates its own assigned chunk, sharing the fitness of the individuals in its chunk with the rest of the processors. Fitness can be encapsulated and broadcasted maximizing the occupation of the underlying packing frames used by the network infrastructure. Moreover, this approach also removes the need for sending the actual rules back and forth between processors as a master/slave approach would require thus, minimizing the communication to the bare minimum the fitness. Figure 6 presents a conceptual scheme of the parallel architecture of `NAX`.

To implement the model presented in Fig. 6, we used `C` and a *message passing interface* (MPI) we used the OpenMPI implementation (Gabriel et al. 2004). Figure 7 shows the code in charge of the parallel evaluation. Each processor computes which individuals are assigned to it. Then it computes the fitness and, finally, it just broadcast the computed fitness. The rest of the process is left untouched, and besides the cooperative evaluation, all the processors end generating the same evolutionary trace.

4.2 Rule sets as individuals

One main characteristic of the so called Pittsburgh style learning classifier systems a particular type of GBML is that individuals encode a rule set (Goldberg 1989; Llorà and Garrell 2001; Goldberg 2002). Thus, evolutionary mechanisms directly recombine one rule set against another one. For classification tasks of moderate complexity, the rule sets are

```
 1. #define VEC_MATCH(vecFLB,fLB,vecFUB,fUB,vecINS,fIN,vecTmp,vecOne,vecRes) {\
 2.     vecFLB = _mm_load_ps(fLB);\
 3.     vecFUB = _mm_load_ps(fUB);\
 4.     vecINS = _mm_load_ps(fIN);\
 5.     \
 6.     vecRes = (__m128i)_mm_cmpgt_ps(vecFUB,vecFLB);\
 7.     vecTmp = _mm_or_si128(\
 8.                     (__m128i)_mm_cmpgt_ps(vecFLB,vecINS),\
 9.                     (__m128i)_mm_cmpgt_ps(vecINS,vecFUB)\
10.                  );\
11.     vecRes = _mm_andnot_si128(_mm_and_si128(vecRes,vecTmp),vecOne);\
12. }
13.
14. void match_rule_set ( RuleSet * rs, InstanceSet is, int iDim, int iRows ) {
15.     int i,j,k,iCnt,iClsIdx,iGround,iPred;
16.     register int iMatcheable;
17.     Instance ins;
18.
19.     __m128i vecRes,vecTmp,vecOne;
20.     __m128 vecFLB,vecFUB,vecINS;
21.
22.     vecOne = (__m128i){-1,-1};
23.
24.     iClsIdx = rs->iCorrectedDim;
25.     clean_fitness_rules_set(rs);
26.     for ( i=0 ; i<iRows ; i++ ) {
27.         // Classify the instance
28.         ins = is[i];
29.         iPred=-1;
30.         for ( j=0 ; iPred==-1 && j<rs->iLen ; j++ ) {
31.             iMatcheable = 1;
32.             for ( iCnt=0,k=j*(rs->iCorrectedDim+VBSIF) ;
33.                   iMatcheable && k<j*(rs->iCorrectedDim+VBSIF)+rs->iDim ;
34.                   k+=VBSIF,iCnt+=VBSIF ) {
35.                 VEC_MATCH(vecFLB,&(rs->pfLB[k]),
36.                           vecFUB,&(rs->pfUB[k]),
37.                           vecINS,&(ins[iCnt]),vecTmp,vecOne,vecRes);
38.                 iMatcheable = 0xFFFF==_mm_movemask_epi8(vecRes);
39.             }
40.         if ( iMatcheable )
41.             iPred = rs->pfLB[j*(rs->iCorrectedDim+VBSIF)+rs->iCorrectedDim];
42.         iPred = (iPred==-1)?rs->iClasses:iPred;
43.         iGround=(int)ins[iClsIdx];
44.         rs->pConfMat[iGround][iPred]++;
45.     }
46. }
```

Fig. 5 This figure presents a vectorized implementation of the rule matching process presented in Fig. 4. Lines 1 12 implement the parallelized test against four attributes using vector instructions. The code is written using C intrinsics for SSE2 compatible architectures. This code runs on P4 or newer Intel processors and Opteron or Athlon 64 AMD processors

not large. However, for complex problems, the potential number of required rules to ensure proper classification may need large amounts of memory that become prohibitive. The requirements increase even further in the presence of noise (Llorà and Goldberg 2003).

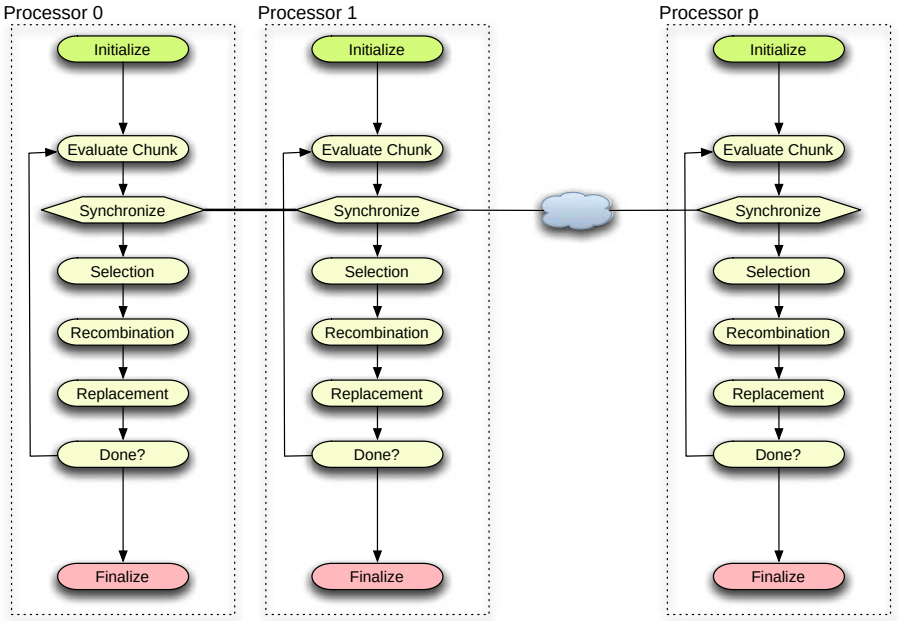


Fig. 6 This figure illustrates the parallel model implemented. Each processor is running the same identical NAX algorithm. They only differ in the portion of the population being evaluated. The population is treated as collection of chunks where each processor evaluates its own assigned chunks sharing the fitness of these individuals with the rest of the processors. This approach minimizes the communication cost

Parallelization may not help much if we need to send large rule sets across the commu nication network. For such reasons, GBML techniques work very well on moderate complexity problems (Bernadó et al. 2001; Bacardit and Butz 2006). However, they need to be modified to deal with complex and large data set, and also avoid the boundaries imposed by the issues mentioned above.

4.3 NAX: Incremental rule learning for very large data sets

An incremental rule learning approach may alleviate memory footprint requirements by evolving only one rule at a time, hence, reducing the memory requirements. However, one rule by itself cannot solve complex problems. For such a reason, each evolved rule is added to the final rule set, and the covered examples are removed from the current training set. The process is repeated until no instances are left in the training set. This approach already introduced by Cordón et al. (2001) and later also used by Bacardit and Krasnogor (2006) allows maintaining relatively small memory footprints, making feasible processing large data sets as the prostate tissue classification data set. However, an incremental approach to the construction of the rule set requires paying special attention to the way rules are evolved. For each run of the genetic algorithm used to evolve a rule, we would like to obtain a maximally general and maximally accurate rule, that is, a rule that covers the maximum number of example without making mistakes (Wilson 1995).

```
 1. void evaluate_population ( Population * pp, InstanceSet is, int iDim, int iRows )
 2. {
 3.    int i;
 4.
 5.    /* Compute the fragments of this processor */
 6.    int iFrag = pp->iLen/FCS_processes;
 7.    int iInit = FCS_process_id*iFrag;
 8.    int iLast = (FCS_process_id+1==FCS_processes)?
 9.                     pp->iLen:
10.                     (FCS_process_id+1)*iFrag;
11.    int iCnt  = 0;
12.    int j,k,l;
13.
14.    /* Create the bucket for the broadcast */
15.    float faFit[2*iFrag];
16.    float faTmp[2*iFrag];
17.
18.    /* Evaluate the given chunk assigned to the processor */
19.    for ( i=iInit,iCnt=0 ; i<iLast ; i++,iCnt++ ) {
20.       match_rule_set(pp->prs[i],is,iDim,iRows );
21.       compute_raw_accuracy_fitness_rule_set(pp->prs[i]);
22.       faFit[iCnt] = pp->prs[i]->fFitness;
23.    }
24.
25.    /* Broadcast each of the chunks */
26.    for ( i=0 ; i<FCS_processes ; i++ ) {
27.       MPI_Bcast((i==FCS_process_id)?faFit:faTmp,iCnt,MPI_FLOAT,i,MPI_COMM_WORLD);
28.       if ( i!=FCS_process_id )
29.          for ( l=0,j=i*iFrag, k=(i+1)*iFrag ; j<k ; j++,l++ )
30.             pp->prs[j]->fFitness = faTmp[l];
31.       }
32. }
```

Fig. 7 This figure presents an implementation of the proposed parallel evaluation scheme using `C` and `MPI`. The piece of code presented below is the only one modified to provide such parallelization capabilities. Each processor computes which individuals are assigned to it (lines 6 10), then it computes the fitness (lines 10 23), and then it just broadcast the computed fitness (lines 26 31)

Llorà et al. (2007) have shown that evolving such rules is possible. In order to promote maximally general and maximally accurate rules à la XCS (Wilson 1995), we compute the *accuracy* (α) and the *error* (ε) of a rule (Llorà et al. 2005). The *accuracy* is the proportion of overall examples correctly classified, and the *error* is the proportion of incorrect clas sifications issued. For simplicity reasons, we use the proportion of correctly issues classifications instead, simplifying the final fitness calculation. Let n_{t+} be the number of positive examples correctly classified, n_{t-} the number of negative examples correctly classified, n_m the number of times a rule has been matched, and n_t the number of examples available. Using these values, the *accuracy* and *error* of a rule r can be computed as:

$$\alpha(r) = \frac{n_{t+}(r) + n_t \ \ (r)}{n_t} \tag{3}$$

$$\varepsilon(r) = \frac{n_{t+}(r)}{n_m(r)} \tag{4}$$

Once the *accuracy* and *error* of a rule are known, the fitness can be computed as follows.

$$f(r) = \alpha(r) \cdot \varepsilon(r)^{\gamma} \tag{5}$$

where γ is the error penalization coefficient. The above fitness measure favors rules with a good classification accuracy and a low error, or maximally general and maximally accurate rules. By increasing γ, we can bias the search towards correct rules. This is an important element because assembling a rule set based on accurate rules guarantees the overall performance of the assembled rule set. In our experiments, we have set γ to 18 to strongly bias the search toward maximally general and maximally accurate rules.

NAX 's efficient implementation of the evolutionary process is based on the techniques described using hardware acceleration Sect. 4.1.1 and coarse grain parallelism Sect. 4.1.2. The genetic algorithm used was a modified version of the *simple genetic algorithm* (Goldberg 1989) using tournament selection (s = 4), one point crossover, and mutation based on generating new random boundary elements.

5 Experiments

This section present the results achieved using NAX. To allow the reader to compare with other techniques, we compare the results obtained using NAX on small data sets provided by the UCI repository (Merz and Murphy 1998) to other well known supervised learning algorithms. Finally, we present the first results on the prostate tissue prediction obtained using NAX. Results focus on the viability of the NAX approach.

5.1 Some UCI repository data sets

The UCI repository (Merz and Murphy 1998) provides several data sets for different machine learning problems. These data sets have been widely used to test traditional machine learning and GBML techniques. Table 1 list the data sets used. Due to the nature of the prostate tissue type classification, we only chose data sets with numeric attributes. Three of these data sets are of relevant interest: (1) son, by far the one with larger dimensionality, (2) gls, the one with large number of classes, (3) tao, proposed by Llorà and Garrell (2001), having complex and non linear boundaries.

Table 1 Summary of the data sets used in the experiments

ID	Data set	Size	Missing values(%)	Numeric attributes	Nominal attributes	Classes
bre	*Wisconsin Breast Cancer*	699	0.3	9		2
bpa	*Bupa Liver Disorders*	345	0.0	6		2
gls	*Glass*	214	0.0	9		6
h s	*Heart Stats Log*	270	0.0	13		2
ion	*Ionosphere*	351	0.0	34		2
irs	*Iris*	150	0.0	4		3
son	*Sonar*	208	0.0	60		2
tao	*Tao*	1888	0.0	2		2
win	*Wine*	178	0.0	13		3

Table 2 Experimental results: percentage of correct classifications and standard deviation from stratified ten fold cross validation runs

ID	0 R	C4.5	NAX
bre	65.52 ± 1.16	95.42 ± 1.69	96.43 ± 1.72
bpa	57.97 ± 1.23	65.70 ± 3.84	64.07 ± 8.36
gls	35.51 ± 4.49	65.89 ± 10.47	68.02 ± 8.69
h s	55.55 ± 0.00	76.30 ± 5.85	75.56 ± 9.39
ion	64.10 ± 1.19	89.74 ± 5.23	89.19 ± 5.27
irs	33.33 ± 0.00	95.33 ± 3.26	94.67 ± 4.98
son	53.37 ± 3.78	71.15 ± 8.54	73.62 ± 9.72
tao	49.79 ± 0.17	95.07 ± 2.11	97.41 ± 0.92
win	39.89 ± 3.22	93.82 ± 2.85	94.34 ± 6.09

Paired *t* test comparisons showed no statistically significant differences between C4.5 and NAX results

0 R result are just provided as guiding base line

We could have chosen complex algorithms as baselines for NAX . However, we would not be able to use them to repeat the experimentation on the prostate tissue classification domain. The algorithms used in the comparison presented in Table 2 were 0 R (Holte 1993) (a simple base line based on majority class classification) and C4.5 (Quinlan 1993). Results show percentage of correct classifications and standard deviation from stratified ten fold cross validation runs. Paired *t* test comparisons showed no statistically significant differences between the pruned tree produced by C4.5 and NAX results. This experiments also helped validate the distributed implementation proposed by NAX. Further results on empirical comparisons can be found elsewhere (Bernadó et al. 2001; Bacardit and Butz 2006).

5.2 Prostate tissue classification

With the previous results at hand, we ran NAX against the prostate tissue classification data set. The original data set is defined by 93 attributes. In this paper, however, we used the reduced version of this data set proposed by (Fernandez et al. 2005) which contains 20 selected attributes out of the 93 available. The dataset is form by 171,314 records. Our goal was to explore how well NAX could generalize over unseen tissue this is the first step to be able to address the cancer prediction problem. The other reason that motivated such experimentation was to achieve similar accuracy results as the ones published earlier by Fernandez et al. (2005) using a modified Bayes technique. If NAX could perform at the same level, we will also obtain a set of rules of interest to the spectroscopist. The inter pretation of the rules will provide insight on how to interpret the models provided by NAX which could not be done with the models early used by Fernandez et al. (2005).

We conducted stratified 10 fold cross validation experiments to measure the general ization capabilities of NAX for this problem. Since the problem was rather small larger data set are being prepared to be run at the supercomputing facilities provided by the National Center for Supercomputing Applications we run the ten fold cross validation runs in a 3GHz dual core Pentium D computer with 4 GB of RAM. NAX took advantage of the hardware support to speedup the matching process and uses two MPI processes to parallelize as introduced in Fig. 6 the evaluation of the overall population. Each fold

took about one hour to complete, with the entire classification lasting less than half a day. We conducted a simple test of adding a second computer with an identical configuration. The overall time for cross validation was reduced to half. Rough estimates which will better measured when larger experiments are conducted on NCSA super computers show that the sequential portion is around 1:1000 for this small data set. Numbers get better as data set increases, which demonstrates that we will be able to process very large data sets and efficiently exploit larger numbers of processors.

We proposed another measure of effectiveness, namely how many records can be processed per second. Using a single processor with the hardware acceleration mechanisms built into NAX, and the evolved rule set formed by 1,028 rules, the average throughput was around 60,000 records per second. For the prostate tissue classification, it took less than three seconds to classify the entire data set. Once the rule set is learnt, the classification problem falls again into the category of embarrassingly parallel problems (Grama et al. 2003). Since no communication is needed, the speedup grows linearly with the number of processors added with the proper rule set replication and data set chunking. Thus, with the dual core box used we where able to just double the throughput (120,000 records per second) by chunking the data set and use both processors.

The previous results show the benefits of hardware acceleration and parallelization, but NAX was also able to achieve very competitive classification accuracy in generalization, correctly classifying 97.09 ± 0.09 of the records (pixels) during the stratified ten fold cross validation. Figure 8 presents the regenerated prostate tissue classification image presented in Fig. 2 using a rule set assembled by NAX. Figure 8a presents the incorrectly classified pixels. Most of the mistakes by the rule set involve similar tissues with few training records available. This trend was also shown elsewhere (Fernandez et al. 2005). C4.5 does not provide any statistically significant improvement (only a marginal, not statistically significant, 0.7%) and provided large decision trees with more than 5,000 leaves not to mention the lack of scalability when compared to NAX.

The rule set assembled by NAX represents an incremental assembling of maximally general and maximally accurate rules. Thus, we can compute how the accuracy of such ensemble improves as new rules are added. Figure 9 presents the overall accuracy as rules are added. It shows an interesting behavior for classifying prostate tissue. Using only 20 rules out of the 1,028 evolved ones, the overall accuracy is 90%, the incorrectly classified 1.3% pixels, and 8.7% were left unclassified. After inspecting the misclassified pixels most of them belongs to borders between tissues and mislabeling arises from the image discretization one pixel containing different tissue types. Table 3 presents the initial four rules that covering 80% of the instances belonging to the two larger tissue types epithelium and fibrous stroma. Such results are relevant, not only for their accuracy, but also because of the insight they provide to the spectroscopist about the problem structure.

6 Conclusions and further work

This paper has presented the initial results achieved in predicting prostate tissue type using GBML techniques. Being able to classify unseen tissue quickly, reliably, and accurately, is the first step towards the creation of CADx systems that may assist a pathologist diagnosing prostate cancer. We have proposed two main efficiency enhancement techniques for GBML exploiting hardware parallelization via vector instructions and coarse grain parallelism via the usage of MPI libraries which allowed us to approach very large data sets. These techniques, together with an incremental genetics based rule learning approach to

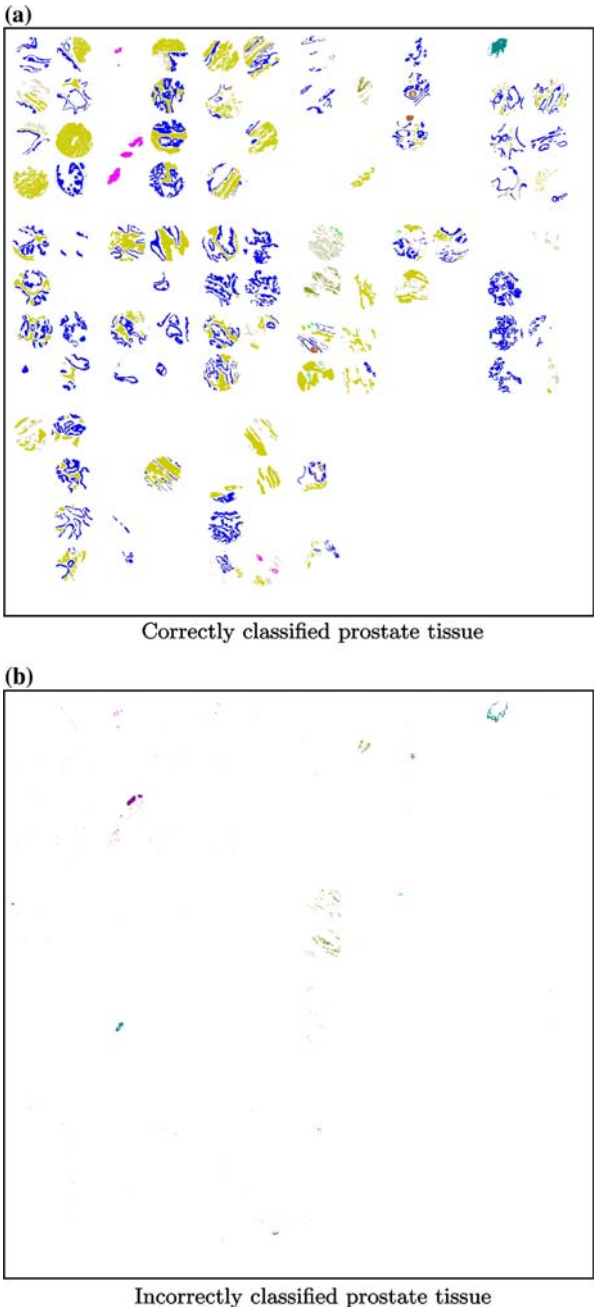


Fig. 8 The figures presented above show the regenerated prostate tissue classification image presented in Fig. 2. (**a**) presents the correctly classified pixels. (**b**) presents the incorrectly classified pixels

assemble rule sets formed by maximally general and maximally accurate rules, have led to the creation of NAX, a system specialized on dealing with large data sets.

Results have shown accurate classification models for prostate tissue along with good scalability of the NAX implementation. Results also reveal peculiarities of the underlying problem structure. With very few rules 20 we were able to correctly classify up to 90%

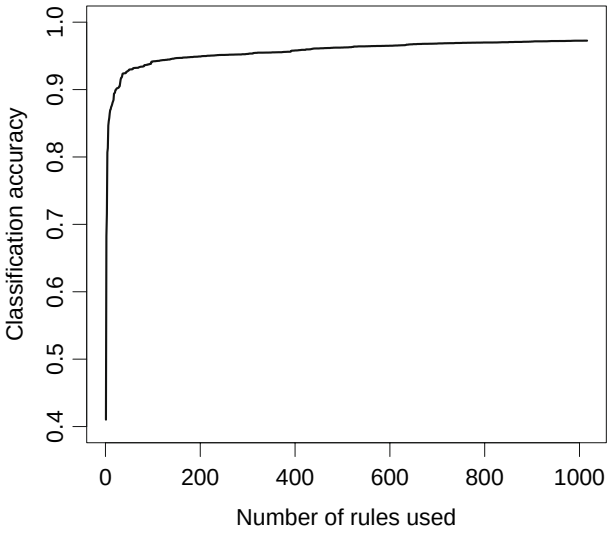


Fig. 9 The rule set as a decision list. The figure presents the classification accuracy as we keep adding rules to the decision list. The first 20 initial rules are able to cover 91% of the records with a classification accuracy of 98.5 90% overall accuracy presented in the figure

Table 3 First top four maximally general and maximally accurate rules that compose the final rule set. The rule set is treated as a decision list, thus we can easily incrementally evaluate the value of the initial four ones

Rule	Rule condition	Tissue type	Accumulated accuracy (%)	Covered records (%)
1.	$0.10 \leq a_1 \leq 0.25$ ^ $0.00 \leq a_4 \leq 0.04$ ^ $1.07 \leq a_8 \leq 2.01$ ^ $0.07 \leq a_{16} \leq 0.16$ ^ $0.25 \leq a_{17} \leq 2.86$ ^ $0.11 \leq a_{18} \leq 0.21$	→ *Fibrous stroma*	41.32	41.96
2.	$0.03 \leq a_1 \leq 0.11$ ^ $0.05 \leq a_7 \leq 0.20$ ^ $1231.88 \leq a_{12} \leq 1247.90$ ^ $1.98 \leq a_{17} \leq 3.83$ ^ $0.13 \leq a_{18} \leq 0.20$	→ *Epithelium*	68.53	69.61
3.	$0.07 \leq a_0 \leq 0.16$ ^ $0.14 \leq a_1 \leq 0.41$ ^ $0.71 \leq a_{10} \leq 1.13$ ^ $1527.54 \leq a_{15} \leq 1533.80$ ^ $0.65 \leq a_{19} \leq 1.50$	→ *Fibrous stroma*	71.59	72.75
4.	$0.05 \leq a_2 \leq 0.09$^ $0.76 \leq a_4 \leq 1.29$^ $1.80 \leq a_6 \leq 2.08$^ $0.17 \leq a_7 \leq 0.24$^ $0.26 \leq a_{16} \leq 0.53$^ $2.79 \leq a_{17} \leq 7.01$^ $0.21 \leq a_{18} \leq 0.32$	→ *Epithelium*	80.78	82.08

of the tissue. Our current work is focused on analyzing the larger data sets containing all the available features and different tissue sources to test the parallelization scalability of NAX on NCSA supercomputers. Once accomplished, the procedure will provide confidence in creating a CADx system to generate a diagnosis based on the evolved models.

Acknowledgments We would like to thank David E. Goldberg for his continual support and encour agement, allowing us to have access to the IlliGAL resources. Thanks also to Kumara Sastry for hallway discussions and to the Automated Learning Group and the Data Intensive Technologies and Applications at the National Center for Supercomputing Applications for hosting this joint collaboration.This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant

FA9550 06 1 0370, the National Science Foundation under grant IIS 02 09199, and the National Institute of Health. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, or the US Government.Rohit Bhargava would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful dis cussions and guidance. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

References

Amdahl G (1967) Validity of the single processor approach to achieving large scale computing capabilities. In Proceedings of the American federation of information processing societies conference (AFIPS). 30:483 485 AFIPS

Bacardit J, Butz M (2006) Advances at the frontier of Learning Classifier Systems. Chapter data mining in Learning Classifier Systems: Comparing XCS with GAssist, vol I. Springer

Bacardit J, Krasnogor N (2006) Biohel: Bioinformatics oriented hierarchical evolutionary learning (Nottingham ePrints). University of Nottingham

Barry A, Drugowitsch J (1997) LCSWeb: the LCS wiki. http://www.lcsweb.cs.bath.ac.uk/

Bernadó E, Llorà X, Garrell J (2001) Advances in Learning Classifier Systems: 4th international workshop (IWLCS 2001). Chapter XCS and GALE: a comparative study of two Learning Classifier Systems with six other learning algorithms on classification tasks. Springer Berlin, Heidelberg, pp 115 132

Bhargava R, Fernandez D, Hewitt S, Levin I (2006) High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. Biochemica et Biophisica Acta 1758(7):830 845

Cantú Paz E (2000) Efficient and accurate parallel genetic algorithms. Kluwer Academic Publishers

Cordón O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems. Evolutionary tuning and learning of fuzzy knowledge bases. World Scientific

Fernandez D, Bhargava R, Hewitt S, Levin I (2005) Infrared spectroscopic imaging for histopathologic recognition. Nat Biotechnol 23(4):469 474

Flockhart I (1995) GA MINER: parallel data mining with hierarchical genetic algorithms (final report). (Technical Report Technical Report EPCCAIKMS GA MINER REPORT 1.0). University of Edinburgh

Gabriel E, Fagg G, Bosilca G, Angskun T, Dongarra J, Squyres J, Sahay V, Kambadur P, Barrett B, Lumsdaine A, Castain R, Daniel D, Graham R, Woodall T (2004) Open MPI: goals, concept, and design of a next generation MPI implementation. In Proceedings of the 11th European PVMMPI Users' group meeting Springer

Goldberg D (1989) Genetic algorithms in search, optimization, and machine learning. Addison Wesley Professional

Goldberg D (2002) The design of innovation: lessons from and for competent genetic algorithms. Springer

Grama A, Gupta A, Karypis G, Kumar V (2003) Introduction to parallel computing. Addison Wesley

Holte R (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11:63 91

Lattouf J B, Saad F (2002) Gleason score on biopsy: is it reliable for predcting the final grade on pathology? BJU Int 90:694 699

Levin I, Bhargava R (2005) Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. Annu Rev Phys Chem 56: 429 474

Llorà X (2002) Genetics based machine learning using fine grained parallelism for data mining. Doctoral dissertation, Enginyeria i Arquitectura La Salle. Ramon Llull University, Barcelona, Catalonia, Euro pean Union

Llorà X (2006) Learning Classifier Systems and other genetics based machine learning Blog. http://www illigal.ge.uiuc.edulcs n gbml/

Llorà X, Garrell J (2001) Knowledge independent data mining with fine grained parallel evolutionary algorithms. In Proceedings of the genetic and evolutionary computation conference (GECCO'2001). Morgan Kaufmann Publishers, pp 461 468

Llorà X, Goldberg D (2003) Bounding the effect of noise in multiobjective Learning Classifier Systems. Evol Comput J 11(3):279 298

Llorà X, Sastry K (2006) Fast rule matching for Learning Classifier Systems via vector instructions. In Proceedings of the 2006 genetic and evolutionary computation conference. ACM Press, pp 1513 1520

Llorà X, Sastry K, Goldberg D (2005) The compact classifier system: motivation, analysis and first results. In Proceedings of the congress on evolutionary computation, vol 1. IEEE press, (Also as IlliGAL TR No 2005019, pp 596 603)

Llorà X, Sastry K, Goldberg D, de la Ossa L (2007) The χ ary extended compact classifier system: linkage learning in Pittsburgh LCS. In Advances at the frontier of Learning Classifier Systems, vol II. IlliGAL report no 2006015. Springer, pp (in preparation)

Merz CJ, Murphy PM (1998) UCI repository for machine learning data bases. http://www.ics.uci.edu/∼mlearn/MLRepository.html

Mitchell T (1997) Machine learning. McGraw Hill

Orriols Puig A, Bernadó Mansilla E (2006) A further look at UCS classifier system. In Proceedings of the 8th annual conference on genetic and evolutionary computation workshop program. ACM Press

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann

Stone C, Bull L (2003) For real! XCS with continuous valued inputs. Evol Comput J 11(3):279 298

Wilson S (1995) Classifier fitness based on accuracy. Evol Comput 3(2):149 175

Wilson S (2000a) Get real! XCS with continuous valued inputs. Lect Notes Comput Sci 1813:209 219

Wilson S (2000b) Mining oblique data with xcs. In Revised papers of the 3th international workshop on Learning Classifier Systems (IWLCS 2000). Springer, pp 158 176

Automated noise reduction for accurate histologic segmentation of tissue from low signal-to-noise ratio spectroscopic imaging data

Rohith Reddy and Rohit Bhargava*
Department of Bioengineering and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

Fourier Transform Infrared (FT-IR) spectroscopic imaging using array detectors often provides low signal-to-noise ratio (SNR) data. One avenue to improving SNR of acquired data is to use a decomposition method to separate high SNR factors from others via a numerical transform e.g. Principal component analysis (PCA). A panel of information-bearing factors could then be selected for inverse transformation to yield high quality data. Selection of factors for inverse transform is usually accomplished manually. In this paper, we propose an approach that utilizes the spatial information in the data set to select factors without human input. An order of magnitude reduction in noise is routinely achievable using this approach. The method is applied to the problem of automating breast tissue histology, in which accuracy in classification of tissue into different cell types is shown to depend on the SNR of data. Using the noise reduction procedure, we were able to recover high classification accuracy with ~ 10-fold lower SNR data. The results imply that ~100-fold reduction in acquisition time is routinely possible for automated tissue classifications by using post-acquisition noise reduction.

***Keywords**:* FT-IR spectroscopy, microspectroscopy, hyperspectral imaging, infrared microscopy, MNF transform, noise reduction, factor selection

1. INTRODUCTION

Fourier Transform Infrared (FT-IR) spectroscopic imaging is a powerful technique to record spatially-resolved chemical information.[1] Large data acquisition rates, as is a typical trade-off for most analytical modalities, lead to degradation in data quality and a consequent loss in the ability to solve problems. Mid-IR sources, interferometers and FPA detectors, further, are such that the overall noise is dominated by detector noise. Following conventional trading rules in IR spectroscopy,[2] hence, the signal is recorded multiple times and added to increase the signal to noise ratio (SNR) of the data. This approach required co-adding a large number of FPA snapshots of the same scene and resulted in long dwell times of the mirror at every optical retardation in initial FT-IR imaging systems.[3] The advantages of this frame co-addition process were limited due to the noise characteristics of the detector. Hence, an optimal combination of frame co-addition and repeated scanning was proposed.[4] The advantages of increasing numbers of scan co-additions were further facilitated by the development of asynchronous[5] and synchronous rapid scan imaging.[6] Fundamentally, these methods all traded the SNR reduction against acquisition time and the trade-off is unavoidable to obtain high SNR data using acquisition-side approaches. Another approach may be to improve hardware but is expensive and impractical for most users. As a consequence, FT-IR imaging is limited in applications that require fast imaging or analysis of large number of samples. For a finite data acquisition time, other schemes to extract low noise information are available[7] but these methods neglect the image as a whole and result in loss of image fidelity. The remaining alternative is to use post-acquisition processing methods.

* To whom correspondence should be addressed. Phone: (217) 265 6596, Fax: (217) 265 0256, email:rxb@illinois.edu

Using computation to enhance instrument performance is becoming an attractive option with the rapid development of powerful computers and increased storage capacities. A procedure based on the Minimum Noise Fraction (MNF) transform,[8] for example, was adopted from the satellite, airborne and other imaging communities[9] for IR spectroscopic imaging.[10,11] Similarly, ideas in data compression and with the potential for attendant noise reduction are being proposed by other groups.[12,13] In this milieu, a general approach to noise reduction is to use an Eigenvalue decomposition of the data using a forward transform, for example, a principal components analysis (PCA). After selecting eigenimages with sufficient SNR, the selected data are inverse transformed to yield the entire dataset with lower noise content. This approach was used[14] to examine phase compositions by enhancing contrast between different regions. PCA reorders data in decreasing order of variance. Similarly, techniques can be used to order eigen images in decreasing order of SNR, which is the aforementioned MNF transform. A modified version[15] of this transform was shown to improve image fidelity and achieve better noise reduction than PCA, for example.

Mathematical transform techniques for noise reduction generally utilize the property that noise is uncorrelated whereas spectra (signals) have a higher degree of correlation. In the transform domain, the signal becomes largely confined to a few eigenvalues whereas the noise is spread across all. Noise reduction can be achieved by retaining eigenvalue images that corresponding to high signal content and computing the inverse transform. To generalize, the images of eigenvalues may be called factors. It is the relative proportion of the signal and noise which forms a criterion for inclusion of specific factors in the inverse transform. Inclusion of too many factors will not allow for significant noise rejection, while inclusion of too few would result in loss of fine spectral features. Hence, identifying eigenvalues corresponding to high signal content is an important step in the noise reduction process. There are many dimension reduction and noise reduction schemes proposed[16,17] Most methods[16,18,19] choose all factors before a certain cut off (*k*) determined based on predefined criteria, thereby placing the ordering burden on the factoring algorithm. However, the assumption that all of the first *k* factors are important is questionable. The MNF approach was specifically developed to overcome the observation that the first *k* factors in PCA were not always optimal. Other methods[17,20] can be computationally expensive or do not utilize some of the features of the data.

Another general criticism of present methods is that they do not explicitly account for the correlated spatial and spectral information in the data. For example, spatial PCA separates features in the spatial domain by accounting for variance in the scene whereas spectral-based PCA may consider a column of spectra without regard to their spatial correlation. The variance in data may arise from the measurement noise, sensor characteristics or may be an artifact. For example, the MNF approach can be shown to rigorously order images in decreasing order of random noise. Implicitly, the signal in the re-ordering of MNF factors is assumed to arise from features in the image but could come from factors other than the sample of interest. We present such a case in Figure 1, which shows the 4^{th}, 8^{th}, 12^{th} and 19^{th} MNF factors for FT-IR data from a breast tissue sample. The 4^{th} MNF factor shows interesting tissue structural features. Although the 8^{th} factor has higher SNR compared to the 12^{th} or 19^{th} factor, the 12^{th} and 19^{th} factors seemingly contain more features of interest. We would include the 12^{th} and 19^{th} factors but not the 8^{th} in a noise reduction scheme. The 8^{th} factor likely arises from illumination or water vapor differences and not from the sample itself. A generalization of the MNF transform has been proposed.[21] However, we did not observe the kind of distortion described in ref. 21 in our data and therefore did not find the need to use the generalized MNF.

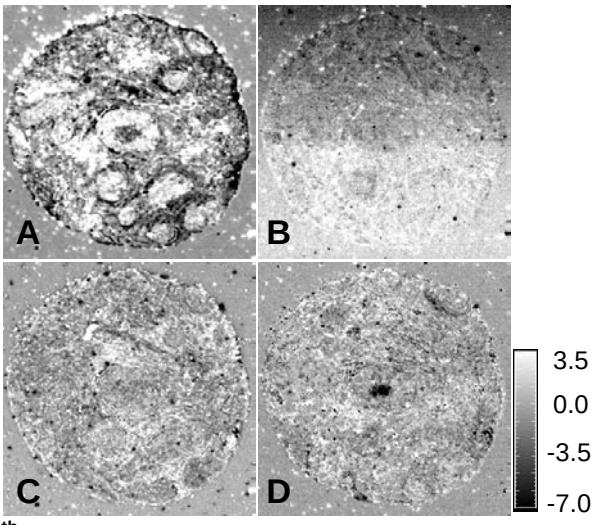


Figure 1. (A) 4th MNF Factor (Tissue features are apparent) (B) 8th MNF factor (C) 12th MNF factor (D) 19th MNF factor. The 8th factor has less apparent structural features than others and is dominated by measurement artifacts.

Given the lack of a generally successful automated approach, the identification of factors to include is invariably a manual process and is the key impediment to routine application of noise rejection methods via factor analysis. First, the manual selection will vary from practitioner to practitioner, leading to potential variance in scientific conclusions or confidence in results. Second, the need to examine every eigenvalue image (or, at least, a large set of images) is time-consuming. The decision to exclude or include images with questionable content is especially difficult and requires significant time as some quantitative guidance is often used. For example, comparisons of values from known sample and sample-less regions may be used to guide the manual selection approach. These two factors are key barriers in the use of post-processing techniques for enhancing IR imaging data.

In this manuscript, we propose a method to automatically determine factors to use in an inverse transform for effective noise rejection. The proposed algorithm selects eigenvalue images based on structural features in a quantitative manner by utilizing both the correlation between spectra as well as the spatial information in the image. We then note that the accuracy of histologic segmentation in breast tissue is decreased as the SNR of data decreases beyond a threshold. Hence, we utilize the automated noise rejection algorithm to recover data quality from low SNR data such that accuracy of tissue recognition is maintained. Last, the improvements in SNR are quantified and discussed in terms of potential data acquisition strategies.

2. METHODS

2.1. Mathematical Background to the Proposed Method

The MNF transform was introduced by Green et. al.[8] to order multispectral data in terms of image quality and we briefly describe the background to our approach next. Consider a three-dimensional (3-D) dataset $X_k(\vec{t})$ where $\vec{t}$ =(*i, j)* represents spatial data coordinates and *k* denotes the spectral element

number. Let the total number of spectral elements in the data be *M*. Let us assume that $X(\vec{t}) = [X_0(\vec{t}), X_1(\vec{t}), X_2(\vec{t}) \ldots\ldots\ldots X_{M-1}(\vec{t})]^T$ can then be written in the form

$$X(\vec{t}) = S(\vec{t}) + N(\vec{t}) \quad (1)$$

Where $S(\vec{t})$ is the signal and $N(\vec{t})$ is additive noise, which are assumed to be uncorrelated. Consequently, the covariances of X, S and N are related through

$$Cov(X) = Cov(S) + Cov(N)$$
$$\Sigma_X = \Sigma_S + \Sigma_N \quad (2)$$

Σ denotes covariance and specifically Σ_S and Σ_N denote covariance of the signal and noise matrices respectively. The noise fraction for the k^h spectral element is defined as

$$F_k = Var(N_k)/Var(X_k) \quad (3)$$

which is the ratio of noise variance to the total variance of that band. The MNF transform is a linear combination of bands

$$Y_k(\vec{t}) = \sum_{m=0}^{M} \alpha_m^k \, X_m(\vec{t}) \quad (4)$$

such that the noise fraction F_k is minimum for $Y_k(\vec{t})$ among all linear transformations orthogonal to $Y_j(\vec{t})$, *j=0, 1, …k.* The vectors $\alpha^k = [\alpha_0^k, \alpha_1^k, \alpha_2^k \ldots\ldots\ldots \alpha_{M-1}^k]^T$ are the left hand eigen vectors of $\Sigma_N \, \Sigma^{-1}{}_X$ and also that the eigen value corresponding to α^k is equal to the noise fraction of Y_k, i.e.

$$\lambda_k = F_k \quad (5)$$

The definition of MNF would imply that $\lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{M-1}$. Since λ_k corresponds to the noise fraction, MNF orders bands in terms of increasing F_k or equivalently, in terms of decreasing SNR. The same set of eigen vectors is obtained from maximizing SNR or the noise fraction. However, the approach that maximizes SNR would result in higher eigen values corresponding to higher SNR and the MNF transform would result in decreasing order of SNR corresponds to decreasing order of eigen values $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{M-1}$. Our implementation uses this approach to compute MNF transforms. It is useful to note that since the MNF transform depends on signal to noise *ratio* it is invariant under scale changes to any band (unlike principal components). It is also useful to note that MNF orthogonailzes $S(\vec{t})$, $N(\vec{t})$ and $X(\vec{t})$.

2.2. Proposed Algorithm based on MNF-transform for Noise Reduction

The MNF transform is computed to obtain factor images corresponding to decreasing SNR values, following the method above. In heterogeneous materials and tissue, we note that the factor images also have structure corresponding to the true structure of the material. The contrast and precise values of signal may not correlate with the spectral image but images having distinct spatial domains will have edges that capture the structural features; this property forms the basis of our factor selection scheme. These features in breast tissue, for example, include boundaries of the sample, ducts and transitions between different structural units. Several methods for edge detection[22] based on different filters and different thresholding schemes have been proposed and studied. Three well known edge detection

techniques (Sobel, Roberts, and Canny) were used and Canny's method[23] was found to be the most effective one for our application. The result of edge detection is a binary image that is termed an 'edge map'. A typical edge map is shown in figure 2. It must be noted that the presence of impulsive noise hinders edge detection. A median filter may be used to mitigate the effect of such impulsive noise. The choice of size of the median filter is a compromise between the size of structural features in the image and the size of noise clusters that need to be removed. Using a large median filter would be effective in removing large clusters of noise but could also result in loss of features, especially those that are smaller than the size of the median filter. Median filters of sizes between 7x7 and 13x13 were found to be most effective for the samples considered here. The edge map in Figure 2, for example, has been obtained after median filtering with a size 9x9 filter.

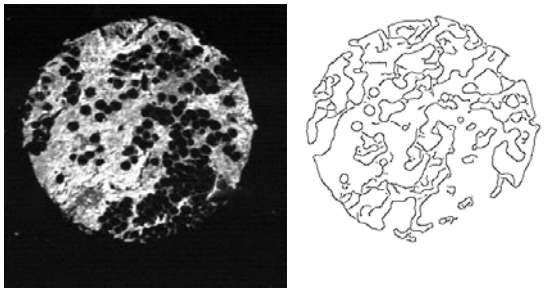


Figure 2. ***Left:*** **Typical 'ideal' image (I)** ***Right:*** **corresponding edge map (E_I)**

The next step in factor selection is to choose an 'ideal', high SNR image (**I**) that has all the structural features of interest. The edge map of **I** and edge maps of factors images are compared to decide whether or not a factor is significant. Since the first MNF factor corresponds to the highest SNR, it could be used as our 'ideal' image **I**. It may also be possible to choose a better image than the first factor in terms of structure if we have some prior knowledge about the sample, for example, information about its spectral characteristics. For many biological tissues, the wavenumber region between ~ 950 cm^{-1} (lower FPA cut-off) to ~1800 cm^{-1} (fingerprint region) and from ~2765 cm^{-1} to ~3750 cm^{-1} (stretching region) is known to have chemical significance. The ideal image **I** could be computed by first calculating the second derivative of spectra in these ranges using a Savitzky-Golay algorithm.[24] The sum of the absolute values of the second derivative data is then indicative of the overall chemical composition of the tissue without regard to scattering artifacts.[25] In general, the fingerprint region of the IR spectrum is likely universally applicable for this procedure. The Savitzky-Golay filter reduces noise while preserving peak heights and widths, and the summation helps improve overall SNR by averaging noise. This gives us a high SNR **I** (figure 2) that captures features from important spectral bands. Yet another alternative is to calculate the Gram-Schmidt intensity of the interferogram of the sample,[26] which could be a faster route by precluding the FT-process. The image, however, would retain both structural and biochemical contributions from all functional groups and scattering interfaces. Yet another approach could be to use the bright field optical microscopy image. The optical image, however, may not contain sufficient contrast, have differences observed in the IR image or may experience a mismatch in resolution. Last, the IR “bright field” equivalent, which is simply the height of the centerburst may be used. Since a background is collected for absorbance data, the sample data set can be easily corrected for illumination differences. This approach can be considered a combination of both IR and visible imaging.

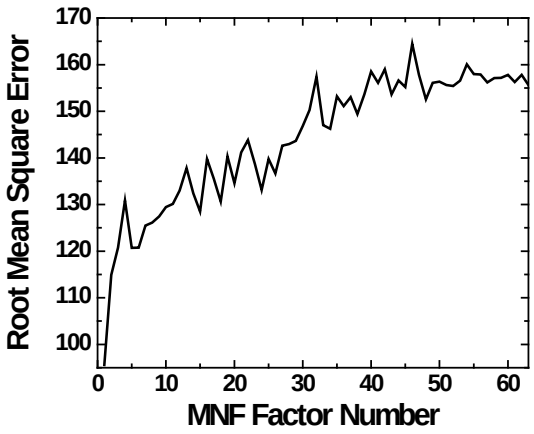


Figure 3. Typical error plot before sorting RMSE

Having chosen an 'ideal' image **I**, its edge map $\mathbf{E}_I$ is computed. Next, each MNF factor image is filtered using the same kernel as that used for the edge map and edge maps $\mathbf{E}_j$, *j=0,1, …,M-1* are found. In practice, the number of significant MNF factors for our data was much smaller (<60) than the number of spectral bands (~1640) and it would be prudent to consider a smaller subset to save computation time. Next, the root mean square error (RMSE) between $\mathbf{E}_I$ and $\mathbf{E}_j$, *j=0,…,M-1* is computed. A typical plot of RMSE vs factor number is shown in Figure 3. RMSE here is an estimate of the spatial similarity of $\mathbf{E}_i$ to **I**. The plot reveals that factors corresponding to higher eigen values may not necessarily have more significant features.

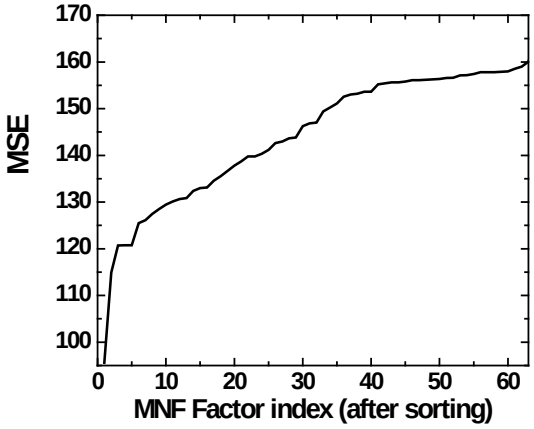


Figure 4. RMSE plot after reordering MNF factors for decreasing spatial similarity to the lowest noise edge map.

A typical RMSE plot after sorting is shown in Figure 4 and can be understood in conjunction with the spatial features in Figure 5. MNF factors and their corresponding edge maps in Figure 5 demonstrate that images with significant features (e.g. Factor 1, 2, 6 and 18) have well defined edge maps while those without significant features (e.g. Factor 44) have nondescript edge maps. The spatial similarity of early factors with the reference edge map results in lower RMSE values that increase with increasing noise. The consistent difference between the reference edge map and edge maps corresponding to noise results in the plateau region of the RMSE curve. Therefore, a good cut-off point for factor selection would be a point on the curve just before the onset of the plateau. Factors close to the chosen cut off in Figure 4 (e.g. Factor 37) have edge maps with a semblance of feature edges buried in noise. By choosing all factors corresponding to RMSE values less than that at the cut-off point, we select only

those factors with significant features. The derivative of the curve in the plateau region is zero and this could be utilized in finding the cut-off point. To mitigate the effect of local variation, a moving average filter may be used to smooth the curve prior to calculating its derivative. The cut-off in our case is chosen to be the point after which the derivative does not rise more than $\mu+3\sigma$ where μ and σ correspond to the mean and standard deviation of the derivative of flat region of the curve. This is a very strict condition which maintains a high degree of spectral detail. Computing the MNF transform, selecting factors based on edge maps and computing the inverse MNF using these factors gives a complete automated noise reduction algorithm that does not require human input. There are choices that can be made while setting up the protocol, for example, in choice of the reference image, that are under operator control. Once the protocol is finalized, however, the process is entirely automated and can be high throughput. Thus, the criteria of both objectivity and automation for noise reduction are addressed.

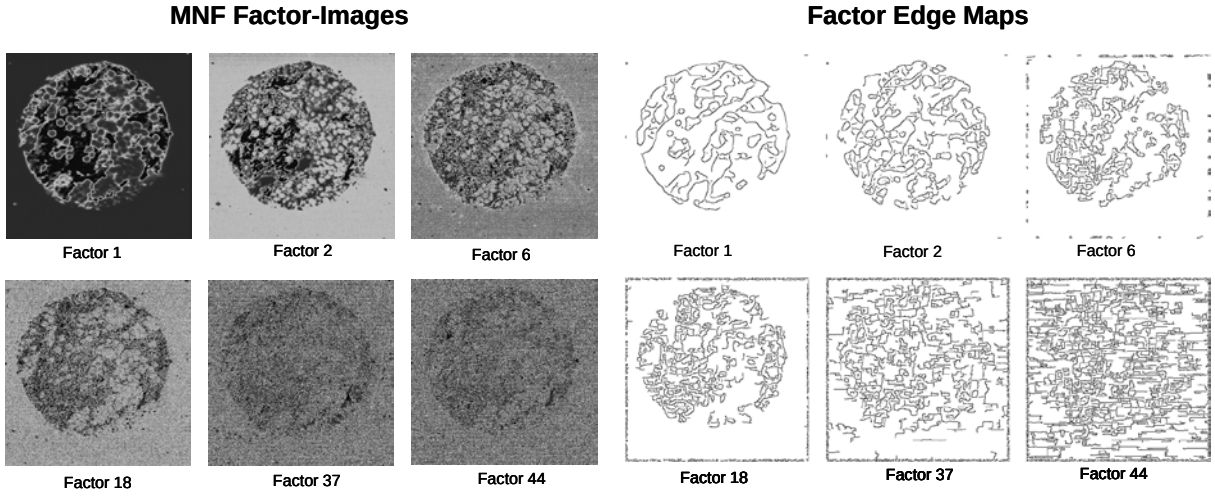


Figure 5. Typical MNF factor images and corresponding edge maps

3. EXPERIMENTAL

Tissue used for this study was obtained from a commercial source (Biomax Inc.) and processed as per procedures reported earlier.[27] Data is acquired at 6.25µm pixel size and a 4 cm^{-1} spectral resolution using the Perkin-Elmer Spotlight 400 imaging spectrometer. A background single beam reference is collected by averaging 120 scans. Sample data sets are acquired by averaging two interferometer scans. An undersampling ratio of two, zero-filling factor of two and N-B medium apodization are employed. To generally validate the method for different instruments, we implemented the same algorithm on data acquired from a system equipped with a larger two-dimensional array detector (Varian Stingray). The system consists of a Varian 7000 Spectrometer coupled to a microscope accessory, UMA-400.The imaging detector is a liquid nitrogen cooled Santa Barbara focal plane FPA of 32 x 32 mercury cadmium telluride (MCT) elements imaging an average spatial area of 175µm x 175µm. The data were acquired in rapid scan mode with an undersampling ratio of 2 at a spectral resolution of 4 cm^{-1} and processed using a factor of two zero-filling and NB-medium apodization. For these data, the number of co-additions was varied (1, 2, 4, 8, 16, 32 and 64 scans) to obtain a range of poor to good SNR data. The background reference was collected at 120 co-additions.

All software used was written in-house or utilized programs in ENVI/IDL. Computing MNF transforms involves estimating noise statistics. ENVI can use a shift difference method to compute noise statistics,

which assumes that every pixel contains both signal and noise, and that adjacent pixels contain the same signal but different noise. A shift difference is performed on the data by differencing adjacent pixel above and to the right of each pixel and averaging the results to obtain the 'noise' value to assign to the pixel being processed. To the extent that this assumption is not true, the noise statistics estimate is in error. Rigorously, the noise should be estimated using repeat measurements, as that is easily possible in FT-IR imaging. With the commercial raster scanning system, however, we were unable to obtain successive measurements without a new scan. The positioning error on the stage was such that slight pixel shifting was observed, hence, precluding true averaging at every pixel. Hence, we considered the shift difference method. The pixel size being set smaller than the lowest resolution achievable, however, and the general nature of large phases in the data here likely result in the estimate being close.

4. RESULTS AND DISCUSSION

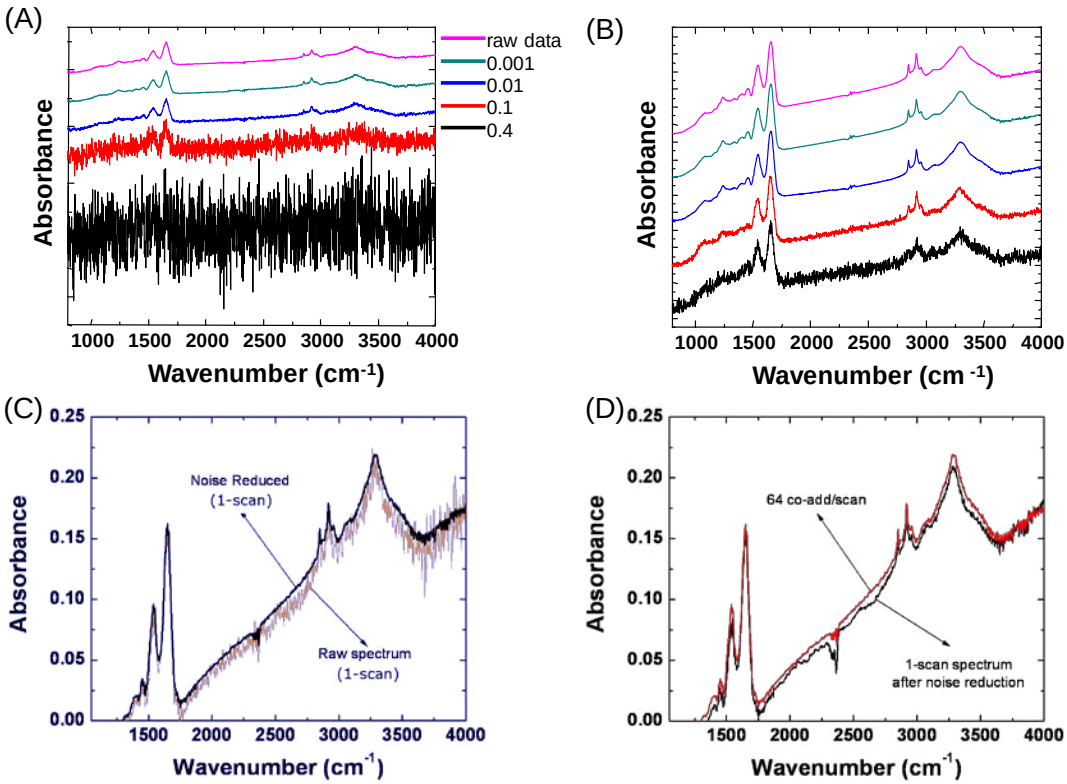


Figure 6 (A) Acquired high SNR data and simulated noisy spectra (σ = 0.001, 0.01, 0.1 and 0.4 a.u.), showing the degradation in data quality Spectra are offset for clarity. (B) Spectra after noise reduction. (C) Absorption spectrum (1-scan) compared to the resulting spectrum from the same pixel after noise reduction. (D) Comparison of 1 scan (noise-reduced) to 64 scans (as-acquired).

4.1. Noise reduction and quantitative benefits

In order to quantify the SNR gain from noise reduction, we first acquired high SNR data using the linear array system as a base for simulations and a comparator. Poor SNR data is simulated from this data by adding noise from a normal distribution with different standard deviations (σ_N = 0.001, 0.01, 0.1 and 0.4 a.u.) as shown for a single pixel in Figure 6(A). Resulting spectra after noise reduction are shown in Figure 6(B). An improvement is apparent, even in cases where noise appears overwhelming. The reduction in

noise achieved is quantified in Figure 7. Noise values were calculated using the non-absorbing 1950 cm^{-1} - 2000cm^{-1} region with 41 spectral points around 1975cm^{-1} and are averages of 1024 spectra.

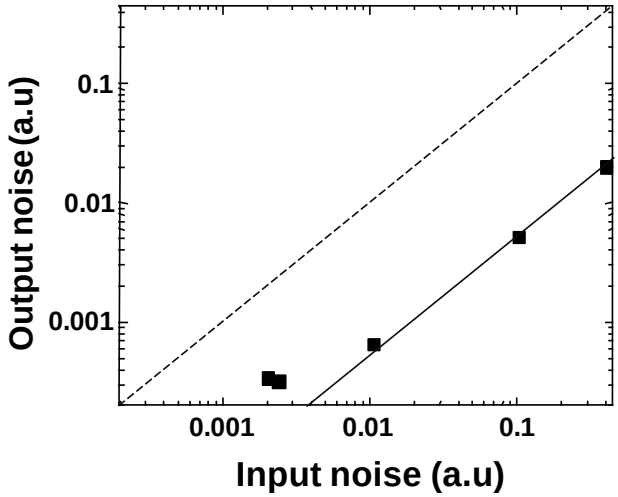


Figure 7 Noise before (input noise) and after application of the algorithm (output noise). An order of magnitude improvement can be observed.

The dashed diagonal is the unity gain line that separates decrease or increase in noise upon application of the algorithm. The plot indicates success in applicability over three orders of magnitude of input noise where an order of magnitude noise reduction is observed. The actual noise reduction depends on the number of factors chosen for the inverse transform, the number of pixels in the original data set and the degree of correlation in the noise. If the noise is high enough, the benefit is observed to be proportional to the input noise. For very low noise cases, the plot indicates that it becomes difficult to improve the data further. This behavior likely arises from the distribution of noise and information in factors. It must be noted that many of the factors rejected in the inverse transform do contain information and all factors selected do contain noise that is both correlated and uncorrelated. Hence, the limitation of the process arises from both correlated noise and the need to balance information content of factor images with the opportunity to reduce noise, We have used a fairly conservative approach to noise reduction in that fewer factors could have been selected, which may also explain the lack of significant improvements when the input noise is low. It is interesting to note that a previous application of the MNF transform[10] also provided a limit to the improvement possible with this approach, but in that of the high noise limit. There, the high input noise data were found to contain a low frequency response in the spectra of inverse transformed data that limited the noise reduction achieved. In summary, the forward-reverse transform approach appears to be bounded in its ability to improve data quality in both the high noise and low noise cases. These limits must be considered when designing data acquisition protocols that take advantage of this post-processing approach.

4.2. Impact

4.2.1. Data acquisition time

From the trading rules of FT-IR spectroscopy[28], a factor of *n* improvement in SNR requires an increase of n^2 in data acquisition time. Hence, a method to increase data acquisition rate without loss in its quality could involve rapid data collection at a low SNR followed by application of numerical techniques for noise reduction. The order of magnitude improvement, as we show above, allows for close to two orders of magnitude reduction in scanning time. To test this hypothesis, we compared noise reduced data from a single interferometer scan with data obtained by averaging 64 scans (Figure 6(D)). Spectra with only one scan, after noise reduction, closely resemble spectra obtained from 64 scans

experimentally. Caution must be exercised, however, in claiming that mathematical techniques provide precisely equivalent data. As can be seen from the spectra, there are some low frequency noise components in the noise–reduced spectrum that were not eliminated.[29]

Automated noise reduction has important implications in areas where data quality cannot be improved by averaging (e.g. kinetics measurements),[30] for low-throughput configurations such as total internal reflection sampling,[31,32,33] where large quantities of data are acquired or where the analyte signal is low. An interesting test case in to perform histopathology without human intervention[34] faster than with current data acquisition protocols. Briefly, FT-IR microspectroscopy combined with pattern recognition tools[35] is rapidly developing as a potential tool for automated structure[36] and disease recognition[37,38,39] within complex tissue by a number of groups.[40,41] Unfortunately, the time to acquire data from large numbers of samples is prohibitive. For example, a recent study[27] reported the quantitative evaluation of classification using large sample and data sets that required many months to acquire. Reducing data acquisition time through automated noise reduction will help reduce time in laboratory studies. When the approach is translated to clinical venues, it will serve to enhance the speeds and throughput of samples. As an example, Figure 8 illustrates the benefits of using automated noise reduction. Prostate tissue is classified into its constituent cell types. Classification is inaccurate for the higher noise case but is recovered when the noise is reduced. The time for data acquisition for this 500 μm x 500 μm image set was reduced from ~ 45 mins to less than 2 mins. While the result demonstrates qualitative agreement between the classified images, we examine next a detailed quantitative assessment of the fidelity of inverse transformed data and the benefits of noise rejection for tissue classification.

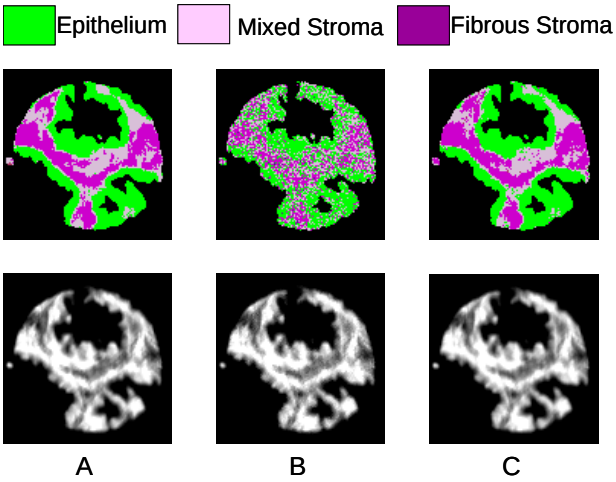


Figure 8. Effect of automated noise reduction on data collection time. Top row: classification results, Bottom row: raw absorbance differences in tissue at 1080 cm^{-1} (A) high SNR data, Noise ~ 0.001a.u. (B) lower SNR data, Noise ~ 0.005a.u. (C) low SNR with noise reduction.

4.2.2. Tissue classification accuracy

Tissue classification accuracy is related to SNR of the data as can be seen in classified images shown in Figure 9. Noise in classified images increases progressively until all ability to segment tissue is lost for noise levels ~0.1 a.u. We quantified classification accuracy, further as measured by calculating the area under the curve (AUC) of the receiver operating curve(ROC)[42] for pixels that meet the threshold for classification, in Figure 9 (E). AUC values finally fall to about 0.5, which is equivalent to random guessing and does not provide any useful classification information. There is a significant decrease in classification accuracy when the noise is greater than 0.1 a.u. in which case some tissue pixels are not even recognized as meeting the threshold for inclusion.

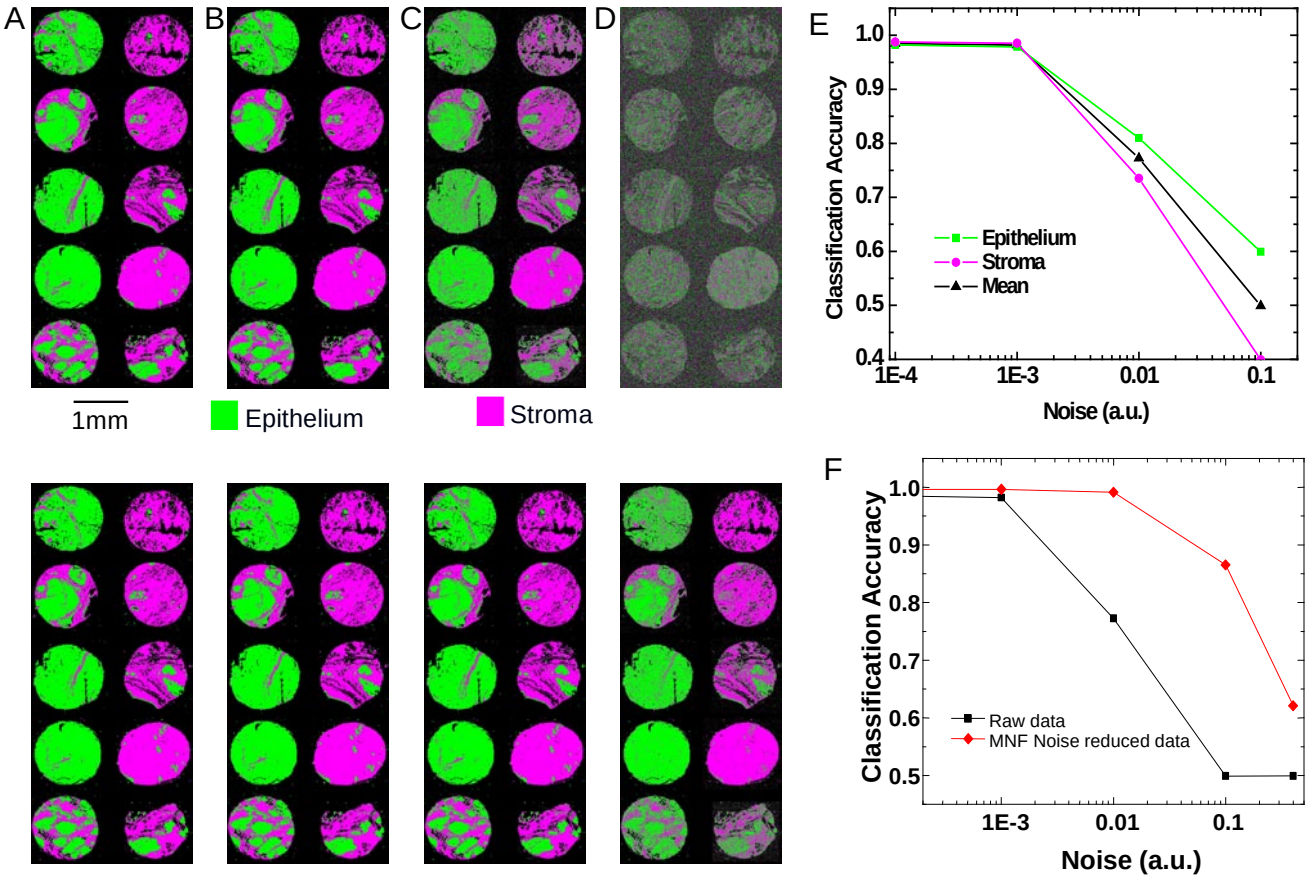


Figure 9. Effect of noise on FT-IR image classification is illustrated for breast tissue in A-D (top panel), where the noise in the data is calculated to be 0.0001, 0.001, 0.001 and 0.1 a.u., respectively. (E) Classification accuracy, as measured by the area under the receiver operating characteristic curve, decreases with increasing noise. Image classification improvement is shown upon using the noise reduction algorithm (A-D, bottom panel). Classified images correspond to noise reduced (A) raw data, (B) 0.001 noise (C) 0.01 noise (D) 0.1 noise (E) Comparing classification before and after noise reduction.

The impact of noise reduction on classification is demonstrated in the bottom panel of Figure 9. Classified images are displayed for each noise-reduced case (A-D) and the classification accuracy values for the noise reduced images are compared with the classification accuracy values for original images (F). Examination of classified images and classification accuracy values indicates that noise reduction scheme improves classifier performance in each case. For as-acquired data and data with noise ~0.001a.u. added, noise reduction does not appear to significantly impact classification since the classification accuracy is almost 100%. On the other hand, noise reduction significantly improves classification from FT-IR spectroscopic imaging data with higher noise levels. Hence, a potential route to faster data acquisition for histopathology, without the need to modify hardware or change any experimental configuration, can be proposed based on post-processing noise reduction. The ten-fold increase in noise of the data to provide the same classification accuracy implies that ~100-fold decrease in data acquisition time may be obtained. Instead of needing ~300 hrs (12 days) to scan a 1 cm x 1 cm area with a large focal plane detector, the proposed approach will allow the same in ~3 hours.

5. CONCLUSIONS

A factor selection scheme based on objective structural features has been proposed here for automated noise reduction after data acquisition. An order of magnitude reduction in noise could be achieved using this algorithm when the noise was not very low. Applied to obtaining results from sample, for example for tissue classification, there is an equivalent recovery of correct results at higher noise levels. The improvement translates directly into a reduction in time required for data collection. It must be noted that the gain here is through post-acquisition computational techniques and does not involve changes in instrumentation hardware or data acquisition schemes. Hence, it is easy to implement and inexpensive to deploy. It is anticipated that the automated nature of the proposed approach will allow it to become routinely applied to enhance data quality and the recover scientific results with lower effort.

6. ACKNOWLEDGEMENTS

The authors thank Frances Nell Pounder for acquiring breast tissue data. Funding for this work was provided in part by the Department of Defense Prostate Cancer Research Program, by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program and Susan G. Komen for the Cure.

7. REFERENCES

(1) Lewis, E.N.; Treado, P.J.; Reeder, R.C.; Story, G.M.; Dowrey, A.E.; Marcott, C.; Levin, I. *Anal. Chem.* **1995**, *67***,** 3377–3381.
(2) Griffiths, P.R. *Anal. Chem.* **1972**, *44*, 1909-1913.
(3) Snively, C.M.; Koenig, J.L.*. Appl. Spectrosc.* **1999,** *53*, 170-177.
(4) Bhargava, R.; Levin, I.W. *Anal. Chem.* **2001**, *73*, 5157 -5167.
(5) Snively, C.M.; Katzenberger, S.; Oskarsdottir, G; Lauterbach. J. *Opt. Lett.* **1999**, *24*,1841–1843.
(6) Huffman, S.W.; Bhargava, R.; Levin, I.W. *Appl. Spectrosc.* **2002**, *56*, 965-969.
(7) Bhargava, R.; Ribar, T.; Koenig, J.L. *Appl. Spectrosc.* **1999**, *53*, 1313–1322.
(8) Green, A.; Berman, M.; Switzer, P.; Craig, M. *IEEE T. Geosci. Remote Sens.* **1988**, *26*, 65–74.
(9) Lee, J. B.; Woodyatt, A. S. ; Berman M. *IEEE T. Geosci. Remote Sens.* **1990**, *28*, 295—304.
(10) Bhargava, R.; Wang, S.; Koenig, J.L. *Appl. Spectrosc.* **2000**, *54,* 486–495.
(11) Wabomba, M.J.; Sulub, Y.; Small G.W. *Appl Spectrosc.* **2007**, 61, 349-358
(12) Vogt, F. *Curr. Anal. Chem.* **2006**, *2*, 107 -127
(13) Vogt, F.; Cramer, J.; Booksh, K. *J. Chemometrics* **2005,** *19*, 510-520
(14) Bhargava, R.; Wang, S.; Koenig, J.L. *Appl. Spectrosc.* **2000**, *54,*1690-1706
(15) Boardman J, Kruse F. *Proc. ERIM Tenth Thematic Conf. Geo. Remote Sens.* **1994**, 407–418.
(16) Wentzell P, Andrews D, Hamilton D, Faber K, Kowalski B. *J. Chemometr.* **1997**, *11*, 339–366.
(17) Qin, S.; Dunia, R. *J. Process Contr.* **2000**, *10*, 245–250.
(18) Cattell R.B. *Multivar. Behav. Res.* **1966**, *1*, 245–276.
(19) Wold S. *Technometrics* **1978**, *20*, 397–405.
(20) Valle, S.; Li, W.; Qin, S*. Ind. Eng. Chem. Res* **1999**, *38*, 4389–4401.
(21) Gordon, C. *IEEE T. Geosci. Remote* **2000**, *38*, 608–610.
(22) Gonzalez, R.; Woods, R.; Eddins, S. *Digital image processing using MATLAB*. Prentice Hall: USA, **2003**, 378-425.
(23) Canny J*. IEEE T. Pattern Anal.* **1986**, *8*, 679–698.
(24) Savitzky, A.; Golay, M. *Anal. Chem.* **1964**, *36*, 1627–1639.
(25) Bhargava, R.; Wang, S.; Koenig, J.L. *Appl. Spectrosc.* **1998**, *52,* 323–328.

(26) Bhargava, R.; Levin, I.W. *Appl. Spectrosc.* **2004**, *58*, 995-1000.
(27) Fernandez, D.C.; Bhargava, R.; Hewitt, S.M.; Levin, I.W. *Nat. Biotechnol*. **2005**, *23*, 469–474.
(28) Griffiths, P.R.; De Haseth, J.A. *Fourier Transform Infrared Spectrometry*. (2nd edn), John Wiley & Sons: New York, USA, 1986, 254-260.
(29) Dongsheng, B.U.; Huffman, S.W.; Seelenbinder, J.A.; Brown, C.W. *Appl. Spectrosc.* **2005**, *59*, 575-583.
(30) Hendershot, R.J.; Fanson, P.T.; Snively, C.M.; Lauterbach, J.A. *Angewandte Chemie* **2003**, *42* 1152-1155.
(31) Sommer, A.J.; Tisinger, L.G.; Marcott, C.; Story, G.M. *Appl. Spectrosc.* **2001**, *55*, 252-256.
(32) Chan, K.L.A.; Kazarian, S.G. *Appl. Spectrosc.* **2003**, *57*, 381-389.
(33) Patterson, B.M.; Havrilla, G.J. *Appl. Spectrosc.* **2006**, *60*, 1256-1266.
(34) Srinivasan G, Bhargava R. Fourier Transform-Infrared Spectroscopic Imaging: The Emerging Evolution from a Microscopy Tool to a Cancer Imaging Modality. *Spectroscopy*, 2007; **22** :30-43.
(35) Lasch P.; Naumann, D. *Cell. Mol. Biol.* **1998**, 44, 189–202.
(36) Mendelsohn, R.; Paschalis, E.P.; Boskey, A.L. *J. Biomed. Opt.* **1999**, *4*, 14-21.
(37) Petibois, C.; Déléris, G. *Trends Biotechnol.* **2006**, *24*, 455-462.
(38) Sahu, R.K.; Argov, S.; Salman, A.; Zelig, U.; Huleihel, M.; Grossman, N.; Gopas, J.; Kapelushnik, J.; Mordechai, S. *J. Biomed. Opt.* **2005,** *10*, 0540171-10.
(39) Krafft, C.; Shapoval, L.; Sobottka, S.B.; Geiger, K.D.; Schackert, G.; Salzer, R. *Biochim. Biophys. Acta – Biomem.* **2006**, *1758*, 883-891.
(40) Diem, M.; Romeo, M.; Boydston-White, S.; Miljkovic´ M.; Matthaus C. *Analyst*, **2004**, 880-885
(41) Diem, M.; Griffiths, P.R.; Chalmers, J.M., eds. *Vibrational Spectroscopy for Medical Diagnosis*, **2008,** John Wiley and Sons.
(42) Hanley, J.A.; McNeil, B.J. *Radiology* **1982,** *143*, 29-36.

Histologic models for optical tomography and spectroscopy of tissues

Rohit Bhargava* and Brynmor J. Davis
Department of Bioengineering and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 North Mathews Ave., Urbana, IL USA 61801

ABSTRACT

Histologic information is often the ground truth against which imaging technology performance is measured. Typically, this information is limited, however, due to the need to excise tissue, stain it and have the tissue section manually reviewed. As a consequence, histologic models of actual tissues are difficult to acquire and are generally prohibitively expensive. Models and phantoms for imaging development, hence, have to be simple and reproducible for concordance between different groups developing the same imaging methods but may not reflect tissue structure. Here, we propose a route to histologic information that does not involve the use of human review nor does it require specialized dyes or stains. We combine mid-infrared Fourier transform infrared (FT-IR) spectroscopy with imaging to record data from tissue sections. Attendant numerical algorithms are used to convert the data to histologic information. Additionally, the biochemical nature of the recorded information can be used to generate contrast for other modalities. We propose that this histologic model and spectroscopic generation of contrast can serve as standard for testing and design aid for tomography and spectroscopy of tissues. We discuss here the biochemical and statistical issues involved in creating histologic models and demonstrate the use of the approach in generating optical coherence tomography (OCT) images of prostate tissue samples.

Keywords: Spectroscopy, histology, Fourier transform infrared spectroscopic imaging, FT-IR, optical coherence tomography, modeling, microscopy, simulation, software phantom, prostate

1. INTRODUCTION

Histopathology is the gold standard for evaluation of microscopic imaging technologies. In particular, correlative information for the evaluation of *in-vivo* imaging technologies usually consists of a corresponding hematoxylin and eosin (H&E) stained image of the excised tissue. Most often, data from an imaging technology are presented side-by-side with the corresponding H&E image. In general, the confirmatory correlation sought is a visual cue that replicates the contrast of H&E images by the particular contrast of the imaging modality. In some cases, contrast information from multiple imaging modalities and dyes can be combined in efforts to reproduce the observed H&E structure. There are few reports, however, on using histopathology as a basis for the development of imaging technologies. In particular, the use of histologic ground truth can help simulate the forward problem and help understand a modality's performance under different experimental parameters, quality of data obtained and potential distortions that may be encountered. Unfortunately, histologic ground truth is not readily available for such use due to the need to stain tissue and the manual nature of examinations.

A new concept has recently emerged in which chemical imaging can be used to measure the intrinsic biochemical content of tissue and employ the same for histologic recognition. The method does not require dyes but relies on spectral data recording. Instead of relying on a human (pathologist) to make decisions, objective numerical algorithms are employed to segment tissue. In one such effort, infrared spectroscopy is used in an imaging format to measure the content of tissue. While efforts to describe tissue using IR spectroscopy are nearly 60 years old, advances in both instrumentation and computational capability have revolutionized this area of investigation [1]. Studies are now being published that involve statistically significant populations of patients (>100), millions of spectral measurements and detailed understanding due to new computational algorithms. The instrumentation advances have mostly centered around the development of Fourier transform infrared (FT-IR) spectroscopic imaging [1, 2, 3].

* rxb@illinois.edu; http://cisl.bioen.illinois.edu

FT-IR imaging combines the imaging capabilities of optical microscopy with the chemical selectivity of FT-IR spectroscopy. Similar to optical microscopy, FTIR imaging of tissue provides images. At each pixel, additionally, is a spectrum that depends uniquely on the sample's chemical composition. IR absorption spectra are, thus, a unique and quantitative "fingerprint" of composition. FTIR spectroscopy is well established for the high-throughput, non-invasive and non-destructive recording of molecular vibrational modes [4]. While not as molecularly specific as some techniques (e.g. mass spectroscopy), spectra provide holistic measurements rapidly and reproducibly [5]. Correspondingly, contrast in FTIR images [6] is generated using algorithms to find differences in tissue chemistry. For example, by selecting appropriate spectral parameters, one is able to "dial" a specific chemistry. Instead of recognizing microscopic structural patterns, we have recently demonstrated an alternate approach to prostate histopathology by directly measuring tissue chemistry [7]. Since both tissue structure and chemical composition are measured, the process is also termed chemical imaging.

A number of laboratory studies have demonstrated carefully biochemical changes indicative of disease [8], determined optical [9] and biological confounding factors [10,11] and demonstrated preliminary potential [12]. Until recently, there was a general lack of substantially validated protocols to perform automated tissue recognition with FT-IR imaging. The primary reason was the slow nature of data acquisition and the trade-off between light intensity and spatial resolution in FT-IR microscopy that prevented validations on large (hundreds) of samples. The drawbacks have been addressed by new technologies and high-throughput sampling methods. For example, a recent report combined FT-IR imaging with combinatorial tissue microarrays (TMA), fast numerical processing and statistical tests [13]. The study developed both objective protocols as well as performed substantial validation [14]. Following the study, protocols to classify tissue into one of ten cell types without human input and without using any stains are available. The net result is that color-coded images of tissue are available that correspond to each cell type. The large population sampling provides an estimate of statistical variance that captures heterogeneity in measurements. The TMA results can be translated to larger, radical prostatectomy (RP) samples or whole mounts without any protocol modifications but few studies have actually demonstrated that result.

In this manuscript, we propose to utilize the advances in automated histologic classification to provide an image that can be used to predict the response of another optical imaging modality. The modality chosen as a demonstrative example is optical coherence tomography (OCT) [15,16,17]. In OCT reflections of low-coherence light are detected permitting the imaging of tissue microstructure in-situ. Micron-scale resolution images can be obtained without the need for excision and histological processing, and the technology has been applied for imaging a wide range of nontransparent tissues [18,19,20,21,22,23]. Imaging depths of 1-2 mm, resolutions of less than 2 μm [24,25] and real-time image acquisition have all been demonstrated [26] using probes [27,28] and for both animal models and humans [29]. The information content of OCT, however, is limited to linear scattering (structure) visualization and will require an expert to interpret data and make decisions. While efforts have been made towards extending OCT for molecular imaging of biological tissue [30,31,32] the gold standard for comparison has generally been H&E images. To simulate the forward problem, the usual approach is to model and validate on spheres suspended in a liquid [33]. Here we show that IR imaging results can be employed to generate computationally the likely image that would be obtained by OCT. A limited set of physical effects is modeled in this preliminary communication but the major idea of using validated histology to simulate image formation in an optical imaging modality is proposed and broadly demonstrated.

2. EXPERIMENTAL

Spectroscopic imaging data sets were acquired for small sample regions separately by averaging two interferometer scans at 4cm^{-1} spectral resolution, and a moving mirror speed of 2.2cm/s on the Perkin-Elmer Spotlight 400 imaging spectrometer with a linear array detector and a raster scanning technique. A NB-medium apodization and undersampling ratio of 2 (referenced to a He-Ne laser) are employed in data collection. As the peaks of interest are the fundamental vibrational modes and the substrate cutoff is approached around 700 cm^{-1}, the free-scanning spectral range is truncated to 4000-720 cm^{-1} for optimal information content and storage. Each pixel in the image recorded was corresponds 6.25μm square at the sample. An IR background is collected with 120 scans per pixel at a location on the substrate with no tissue present and the ratio of the background to sample intensity is computed. Any remaining vapor artifacts are removed from the spectral data using the Perkin-Elmer Spotlight atmospheric correction algorithm. A formalin-fixed, paraffin-embedded radical prostatectomy sample was obtained from an anonymized donor in the NIH tissue array research program. A thin (~5 μm) section of the tissue was microtomed and placed onto a BaF_2 disk. The sample was washed in

hexane for 24 hours and paraffin was removed. Data were processed following previously described procedures [34]. Figure 1 shows the class image obtained after automated tissue segmentation. Figure 2 shows the distribution of protein in tissue based on the Amide I vibrational mode, which is roughly indicative of density differences between tissues.

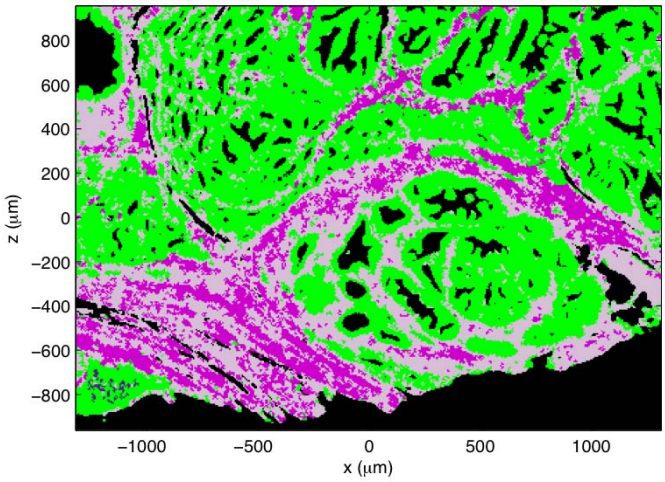


Fig. 1. Classified image of a section of prostate tissue in which colors denote specific cell types. Epithelium, fibroblast-rich stroma and largely extra-cellular matrix classes are denoted by green, magenta and thistle colors respectively.

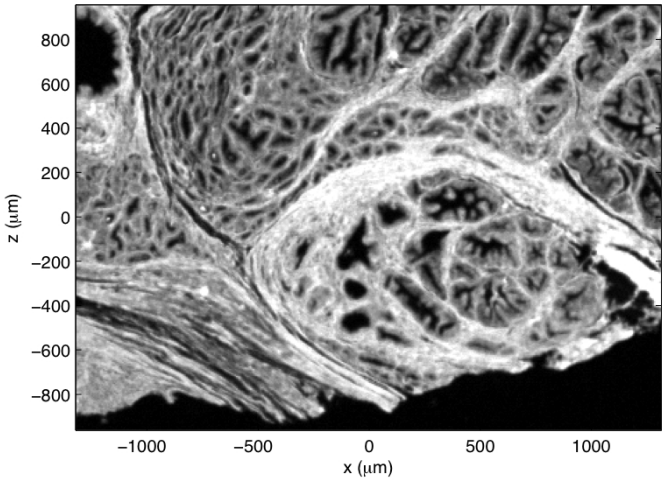


Fig. 2. Grayscale image demonstrating the relative distribution of the Amide I vibrational mode absorbance within tissue. The texture in the image is roughly indicative of the density changes in the structure of the tissue.

3. MODELING

3.1 Transformations between tissue properties

While different imaging techniques operate using different contrast mechanisms, e.g., scattering, absorption, fluorescence, etc., it is generally assumed that contrast is directly correlated to object structure. Consequently, one can reasonably produce mappings between tissue properties using relatively simple, local, mathematical operations. For example, each of the tissue types in Fig. 1 can be modeled as having a different refractive index in the near infrared. Defining an arbitrary unit of index contrast Δn, and assigning refractive indices $(1+\Delta n)$ to epithelium, $(1+1.6\Delta n)$ to

extra-cellular matrix, and $(1+2.4\Delta n)$ to fibroblast-rich stroma, results in the index map seen in Fig. 3. Alternatively, one can directly associate the grayscale values in Fig. 2 with the refractive index contrast.

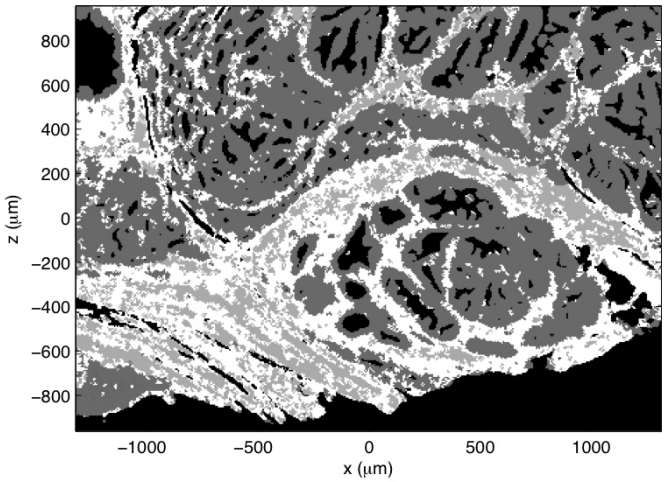


Fig. 3. A near-infrared index contrast or scattering potential (Δn) image calculated by assigning a different refractive index to each class in Fig. 1.

3.2 Simple simulation of OCT image formation

Starting with the maps of scattering potential based on Fig. 2 or as shown in Fig. 3, it is possible to simulate OCT imaging. OCT is a coherent imaging technique that uses low-coherence interferometry to resolve sample structure in the axial (depth) direction and scanning of the focus to resolve transverse structure. Modern OCT systems often collect data in the spectral domain, as spectroscopic data collection offers significant advantages in terms of speed and signal strength [35,36]. Mathematically, two-dimensional spectral OCT data, *S*, can be expressed [37] as

$$S(x,k) = A(k)\int g^2(x,z,k)\Delta n(x,z)dz \ , \qquad (1)$$

where *k* is wavenumber (the spectral axis), *A* is a factor dependent on the spectral profile of the optical source and *g* is the focused field produced by the objective lens. It should be noted that this model employs the first Born approximation, i.e., only singly scattered light is considered. Multiply scattered light does not contribute to the usable image in OCT and results in image artifacts.

In the simple model used here, the illumination source is assumed to have a flat emission profile over the collected bandpass and a Gaussian beam [38] describes the focused field. An objective numerical aperture of 0.025 is simulated and the spectrometer is modeled as collecting 160 data points over a spectral range corresponding to wavelengths between 980nm and 1020nm. With these parameters the transverse resolution of the system is approximately 25.5μm (calculated from the Gaussian beam waist) and the depth of focus is approximately 1020μm (calculated from the Rayleigh range).

To a first approximation, the relation described in Eq. (1) can be inverted by taking an inverse Fourier transform over the *k* axis. The axial resolution of the system is defined by the bandwidth of the source and for the parameters simulated here, this gives a value of approximately 12.5μm. Equation (1) was evaluated for the scattering potentials shown in both Fig. 2 and Fig. 3, and OCT images calculated – the image from Fig. 2 is shown in Fig. 4, and the image from Fig. 3 is shown in Fig. 5. The difference between the images illustrates the basis of contrast. If simple density changes are considered, Fig. 4 provides an image in which the contrast between the classes is limited (cf. Fig. 1). In the case where properties of cell types are explicitly considered, the contrast between classes is accurately reflected in the image.

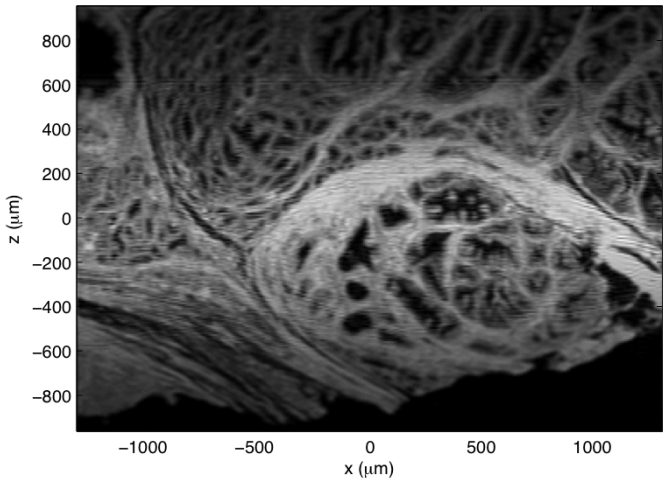


Fig. 4. A simulated OCT image calculated from the scattering potential of Fig. 2. The axial dimension is in z and the transverse dimension is x. The focus of the objective lens is at the z=0 plane and the probing light is assumed to be incident from the bottom of the image.

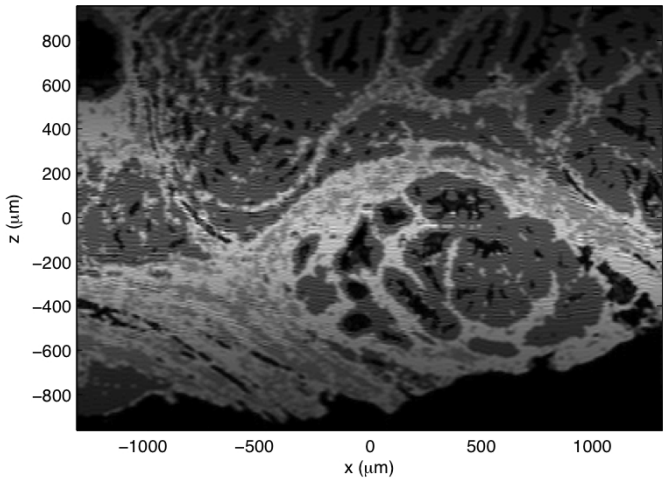


Fig. 5. A simulated OCT image calculated from the scattering potential of Fig. 3.

The effects of the limited OCT resolution can be seen in both images. Also visible is the loss signal strength as one moves outside of the depth of focus. Standard OCT image reconstruction breaks down outside of the depth of focus, resulting in transverse blurring and a rapid decay in signal strength. However, more advanced image reconstruction techniques can be applied to eliminate the transverse blurring and lessen the signal decay [33,39] – these methods could be applied to the simulated data but are not considered here.

3.3 Including noise and absorption

Absorption and multiple scattering limit the depth of imaging achievable in OCT. These effects are not captured in the linearized model of Eq. (1), as applying the first Born approximation involves assuming that the illuminating field at a point is not appreciably affected by the rest of the sample. An exact accounting of absorption and multiple scattering can be difficult, since both effects are non-local, i.e., the absorption of the field to a given point depends on the entire propagation path of the incident light, and the strength of the multiply scattered field at a point depends on the scattering from all neighboring points. However, as a first approximation, the sample absorption can be modeled with a homogeneous absorption coefficient. For a well-collimated beam the incident field will then decay exponentially into the sample. The incident field in Eq. (1) is then replaced by

$$g(x,z,k) \rightarrow g(x,z,k)\exp(-\alpha k z), \qquad (2)$$

where α is the absorption coefficient. This modified forward model was applied to the index map seen in Fig. 3 and the corresponding OCT images are displayed in Fig. 6. Three different absorption coefficients were simulated and it can be seen that the absorption does indeed limit the depth of imaging in these simulations.

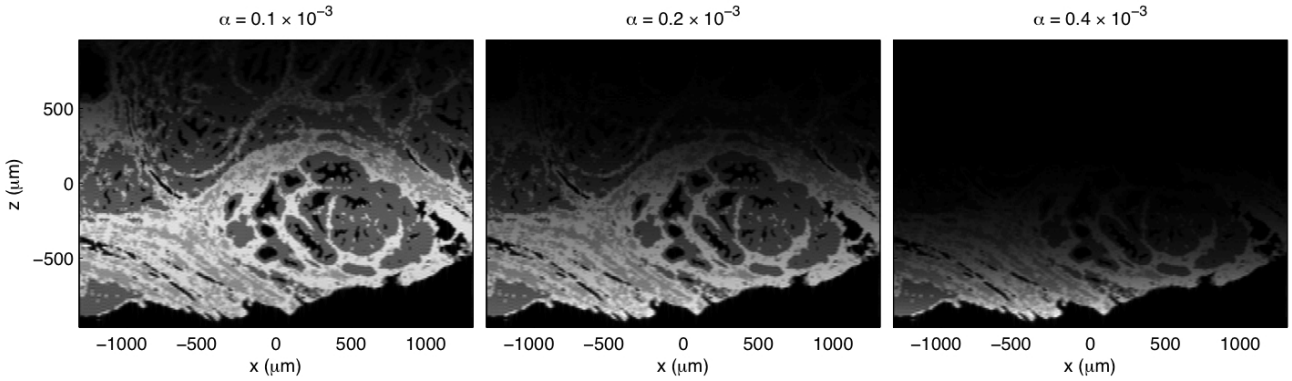


Fig. 6. Simulated OCT images calculated with a forward model that includes absorption. Three different absorption coefficients are simulated, resulting in three differing penetration depths for the illuminating light.

Noise can also be a crucial limiting factor in the performance of OCT imaging systems. Fluctuations of the interferometer reference beam often dominate the statistics, and so noise can be reasonably modeled as Gaussian and spatially independent in the image domain (see Ref. 36 for a more detailed OCT noise model). The effects of noise on the image are illustrated in Fig. 7. In this figure the signal to noise ratio (SNR) is defined as $20\log_{10}(P/\sigma)$, where P is the peak value of the image and σ is the standard deviation of the Gaussian noise.

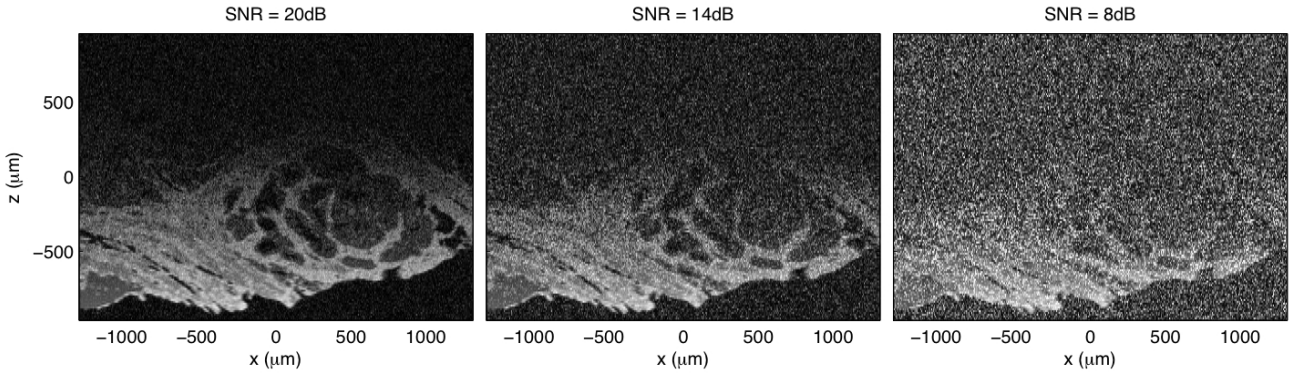


Fig. 7. The $\alpha=0.2\times10^{-3}$ image of Fig. 6 corrupted by varying amounts of measurement noise. All of these images are displayed on the same scale with some high-noise, high-signal pixels saturating the color map.

4. CONCLUSIONS

The use of histologic models derived from spectroscopic imaging and automated analysis of clinical samples has been proposed in this manuscript. While the technology to extract histologic ground truth is becoming readily available and used by many practitioners, the use of such information to model and understand other imaging systems has not been demonstrated. A simple set of image formation steps and effects of experimental parameters were implemented in this manuscript to simulate OCT images from the corresponding class images and spectral texture. The effects of using both approaches, idealized sample absorption and noise can be seen in the generated images. It is anticipated that the methodology proposed here will be refined and may prove useful in the modeling, design and evaluation of imaging instruments for specific tissue types.

ACKNOWLEDGEMENTS

This study was supported in part by a DoD Young Investigator Award in the prostate cancer research program and a grant from the Grainger Foundation. We gratefully acknowledge the prostate tissue samples and many discussions with Dr. Stephen M. Hewitt, National Cancer Institute.

REFERENCES

[1] The state of the art in this field is summarized in a recent volume, Bhargava, R. and Levin, I. W., eds., [Spectrochemical Analysis Using Infrared Multichannel Detectors], Blackwell Publishing, Oxford, (2005).

[2] Lewis, E. N., Treado, P. J., Reeder, R. C., Story, G. M., Dowrey, A. E., Marcott, C. and Levin, I. W., "Fourier transform spectroscopic imaging using an infrared focal-plane array detector," Anal. Chem. 67, 3377-3384 (1995).

[3] Bhargava, R. and Levin, I. W., "Fourier transform infrared imaging: theory and practice," Anal. Chem. 73, 5157-5167 (2001).

[4] Diem, M., Romeo, M., Boydston-White, S., Miljkovic, M. and Matthaus, C., "A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004)," Analyst 129, 880-885 (2004).

[5] Many issues are summarized in Jackson, M., "From biomolecules to biodiagnostics: spectroscopy does it all," Faraday Discuss. 126, 1-18 (2004).

[6] Koenig, J. L., Wang, S. Q. and Bhargava, R., "FTIR images," Anal. Chem. 73, 360A-369A (2001).

[7] Fernandez, D. C., Bhargava, R., Hewitt, S. M. and Levin, I. W., "Infrared spectroscopic imaging for histopathologic recognition," Nat Biotechnol. 23, 469-474 (2005).

[8] McIntosh, L. M., Jackson, M., Mantsch, H. H., Stranc, M. F., Pilavdzic, D. and Crowson, A. N., "Infrared spectra of basal cell carcinomas are distinct from non-tumor-bearing skin components," J. Invest. Dermatol. 112, 951-956 (1999).

[9] Mohlenhoff, B., Romeo, M., Diem, M. and Wood, B. R., "Mie-type scattering and non-Beer-Lambert absorption behavior of human cells in infrared microspectroscopy," Biophys. J. 88, 3635-3640 (2005).

[10] Shaw, R. A., Guijon, F. B., Paraskevas, M., Ying, S. L. and Mantsch, H. H., "Infrared spectroscopy of exfoliated cervical cell specimens – proceed with caution," Anal. Quant. Cytol. 21**,** 292-302 (1999).

[11] E.g. the effect of cell turnover is characterized by Diem et al. (e.g. Boydston-White, S., Gopen, T., Houser, S., Bargonetti, J. and Diem, M., "Infrared spectroscopy of human tissue. V. Infrared spectroscopic studies of myeloid leukemia (ML-1) cells at different phases of the cell cycle," Biospectroscopy 5, 219-227 (1999)) in a number of articles and demonstrated quantitatively in a recent study that has implications for disease diagnoses Mourant, J. R., Yamada, Y. R., Carpenter, S., Dominique, L. R. and Freyer, J. P., "FTIR spectroscopy demonstrates biochemical differences in mammalian cell cultures at different growth stages," Biophys. J. 85, 1938–1947 (2003).

[12] Andrus, P. G., "Cancer monitoring by FTIR spectroscopy," Tech. Cancer Treat. Res. 5, 157-167 (2006).

[13] Bhargava, R., "Towards a practical Fourier transform infrared chemical imaging protocol for cancer pathology," Anal. Bioanal. Chem. 389, 1155-1169 (2007).

[14] Levin, I. W. and Bhargava, R., "Fourier transform infrared vibrational spectroscopic imaging: integration of microscopy and molecular recognition," Annu. Rev. Phys. Chem. 56, 429-474 (2005).

[15] Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A. and Fujimoto, J. G., "Optical coherence tomography," Science 254,1178-1181 (1991).

[16] Fercher, A. F., Drexler, W., Hitzenberger, C. K. and Lasser, T., "Optical coherence tomography – principles and applications," Rep. Prog. Phys. 66, 239-303 (2003).

[17] Schmitt, J. M., "Optical coherence tomography (OCT): a review," IEEE J. Sel. Top. Quant. Elec. 5, 1205-1215 (1999).

[18] Schmitt, J. M., Knuttel, A., Yadlowsky, M. and Eckhaus, A. A., "Optical coherence tomography of a dense tissue: statistics of attenuation and backscattering," Phys. Med. Biol. 39, 1705-1720 (1994).

[19] Schmitt, J. M., Yadlowsky, M. J. and Bonner, R. F., "Subsurface imaging of living skin with optical coherence microscopy," Dermatology 191, 93-98 (1995).

[20] Sergeev, A. M., Gelikonov, V. M., Gelikonov, G. V., Feldchtein, F. I., Kuranov, R. V., Gladkova, N. D., Shakhova, N. M., Snopova, L. B., Shakov, A. V., Kuznetzova, I. A., Denisenko, A. N., Pochinko, V. V., Chumakov, Y. P. and Streltzova, O. S., "*In vivo* endoscopic OCT imaging of precancer and cancer states of human mucosa," Opt. Express 1, 432-440 (1997).
[21] Tearney, G. J., Brezinski, M. E., Boppart, S. A., Bouma, B. E., Weissman, N., Southern, J. F., Swanson, E. A. and Fujimoto, J. G., "Catheter-based optical imaging of a human coronary artery," Circulation 94, 3013 (1996).
[22] Tearney, G. J., Brezinski, M. E., Southern, J. F., Bouma, B. E., Boppart, S. A. and Fujimoto, J. G., "Optical biopsy in human gastrointestinal tissue using optical coherence tomography," Amer. J. Gastroenterol. 92, 1800-1804 (1997).
[23] Tearney, G. J., Brezinski, M. E., Southern, J. F., Bouma, B. E., Boppart, S. A. and Fujimoto, J. G., "Optical biopsy in human urologic tissue using optical coherence tomography," J. Urol. 157, 1915-1919 (1997).
[24] Bouma, B. E., Tearney, G. J., Boppart, S. A., Hee, M. R., Brezinski, M. E. and Fujimoto, J. G., "High resolution optical coherence tomographic imaging using a modelocked Ti:Al203 laser," Opt. Lett. 20, 1486-1488 (1995).
[25] Drexler, W., Morgner, U., Kartner, F. X., Pitris, C., Boppart, S. A., Li, X., Ippen, E. P. and Fujimoto, J. G., "In vivo ultrahigh resolution optical coherence tomography," Opt. Lett. 24, 1221-1223 (1999).
[26] Tearney, G. J., Bouma, B. E., Boppart, S. A., Golubovic, B., Swanson, E. A. and Fujimoto, J. G., "Rapid acquisition of in vivo biological images using optical coherence tomography," Opt. Lett. 21, 1408-1410 (1996).
[27] Tearney, G. J., Boppart, S. A., Bouma, B. E., Brezinski, M. E., Weissman, N. J., Southern, J. F. and Fujimoto, J. G., "Scanning single-mode fiber optic catheter-endoscope for optical coherence tomography," Opt. Lett. 21, 1-3 (1996).
[28] Boppart, S. A., Bouma, B. E., Pitris, C., Tearney, G. J. and Fujimoto, J. G., "Forward-imaging instruments for optical coherence tomography," Opt. Lett. 22, 1618-1620 (1997).
[29] Tearney, G. J., Brezinski, M. E., Bouma, B. E., Boppart, S. A., Pitris, C., Southern, J. F. and Fujimoto, J. G., "In vivo endoscopic optical biopsy with optical coherence tomography," Science 276, 2037-2039 (1997).
[30] Marks, D. L. and Boppart, S. A., "Nonlinear interferometric vibrational imaging," Phys. Rev. Lett. 92, 123905-1-4 (2004).
[31] Vinegoni, C., Bredfeldt, J. S., Marks, D. L. and Boppart, S. A., "Nonlinear optical contrast enhancement for optical coherence tomography," Opt. Express 12, 331-341 (2004).
[32] Applegate, B. E., Yang, C. and Izatt, J. A., "Theoretical comparison of the sensitivity of molecular contrast optical coherence tomography techniques," Opt. Express 13, 8146-8163 (2005).
[33] Ralston, T. S., Marks, D. L., Carney, P. S. and Boppart, S. A., "Interferometric synthetic aperture microscopy," Nat. Phys. 3, 129-134 (2007).
[34] Bhargava, R., Fernandez, D. C., Hewitt, S. M. and Levin, I. W., "High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data," Biochim Biophys Acta. 1758, 830-845 (2006).
[35] Choma, M. A., Sarunic, M. V., Yang, C. and Izatt, J., "Sensitivity advantage of swept source and Fourier domain optical coherence tomography," Opt. Express 11, 2183-2189 (2003).
[36] Leitgeb, R., Hitzenberger, C. K. and Fercher, A. F., "Performance of Fourier domain vs. time domain optical coherence tomography," Opt. Express 11, 889-894 (2003).
[37] Davis, B. J., Schlachter, S. C., Marks, D. L., Ralston, T. S., Boppart, S. A. and Carney, P. S., "Nonparaxial vector-field modeling of optical coherence tomography and interferometric synthetic aperture microscopy," J. Opt. Soc. Am. A 24, 2527-2542 (2007).
[38] Saleh, B. E. A. and Teich, M. C., [Fundamentals of Photonics], Wiley Series in Pure and Applied Optics, (1991).
[39] Davis, B. J., Marks, D. L., Ralston, T. S., Carney, P. S. and Boppart, S. A., "Interferometric synthetic aperture microscopy: computed imaging for scanned coherent microscopy," Sensors 8, 3903-3931 (2008).